OR 2015 Fedora Technical Working Group Panel Proposal

Last year's production release of Fedora 4.0.0 (F4) came with significant capability improvements from previous Fedora versions. However, given the newness of the internal application stack, F4 also came with a variety of questions and concerns. It was not clear to many Fedora stakeholders whether F4 held fast to the project's historic values of preservation and durability, and if the future-facing features performed as expected. It was out of these questions that the Fedora Leadership team formed a working group chartered to analyze and assess the current state of F4 and its suitability as a platform to meet the community's future needs.

This team, termed the Fedora Technical Working Group, established nine areas of which the top four were selected for an in-depth assessment. The areas of assessment were as follows:

- HTTP API
- Clustering
- Preservation-worthiness
- Performance

An assessment was produced for each of these four areas which then resulted in specific outcomes and recommendations.

HTTP API

The view that most users and integration applications have of F4 is of its HTTP API. It is therefore vital that this API not only be well-defined, but more importantly integrate with existing web standards. This area of assessment reviewed the then-extant HTTP API and defined a path to a vision of Fedora as an architecture of standardized APIs.

The most important outcomes from assessment of the HTTP API were:

- Commitments to specify it in an implementation-independent fashion, to partition it into useful modules arranged in intelligible packages, and to provide test packages for each module to enable implementations to guarantee compliance to their audiences.
- A commitment to the W3C's Linked Data Platform specification as the core module of the HTTP API, as well as a commitment for the use of RDF as the syntax for API messages in all modules.
- A commitment to create useful and reusable ontologies expressing the semantics of API messages.

This approach will support a more interoperable and loosely-coupled ecosystem of Fedora services as well as create opportunities for intelligent reimplementation of canonical Fedora services as that work becomes attractive for technological or other reasons. Additionally, the publication of stable ontologies for the API will support durability in Fedora repository

implementations in the way that content modeling has so done for Fedora repository contents over the past fifteen years.

Clustering

Clustering is a repository feature that can potentially address a number of use cases ranging from geographic distribution to high-availability. F4 exposes an initial clustering capability, the boundaries of which the technical working group explored.

The focus of the team's assessment was on read-only and read-dominant workloads over replicated F4 clusters. In other words, the configuration was for a repository that has already been populated or that has moderate additions while supporting replication and high read load. It was demonstrated that when the number of Fedora instances in a cluster increases, the maximum read load the cluster can handle increases proportionally. Read latency in this case does not change significantly, although the write latency showed signs of increase with increasing number of replicated Fedora instances in the cluster. The results give confidence in using replicated F4 clustering to accommodate higher read workload.

Preservation-worthiness

Fedora has a historic foundation as a preservation platform. With the re-architecture in the F4 effort and the new layers in Fedora's underlying component stack, questions surfaced around whether the same preservation capabilities continue to exist in the application's latest incarnation. The investigations explored by the technical working group focused on the following preservation-related activities:

- Verifying that fixity checking could detect transmission and storage errors.
- Testing backup and restore functionality for whole-repository backups.
- Automatically serializing metadata updates to disk with an event-driven message-consumer.
- Testing transparent storage options, including using filesystem federation to serve the contents of a local filesystem through the repository, and verifying that content files on disk are recoverable with a minimal toolset.

Performance

The fourth area of assessment was performance. With the variety of usage scenarios, hardware configurations, and application tuning variables, it is always a challenge to develop a reliable benchmark. Due to the potential for misleading results, all performance testing was rooted in collection sizes and performance expectations from existing and projected repository workflows provided by the institutions of which working group members represented. The approach developed by the technical working group was to define a comprehensive set of usage patterns, design tests that demonstrate those patterns at increasing levels of stress,

and implement a repeatable framework for re-running the tests over time. The team defined the scenarios and related tests, leaving the actual execution as a future task.

Conclusion

F4 has been three years in the coming. The first production version of the repository was released in November of 2014. There were naturally open questions as to whether this major iteration on the Fedora platform addresses historic and future requirements, and where not, if it is architected to eventually close those gaps. A group of vested repository stakeholders formed this technical working group to dig below the surface to understand the truth of F4's capabilities and potentials. This panel will be an open discussion with those who have taken a firsthand, objective investigation of the state of Fedora 4.