Member Nodes - Task #5455

MNDeployment # 3683 (New): Purdue University Libraries

Contact Line Pouchard to re-open discussion about PURL as a potential MN

2014-05-29 19:34 - Bruce Wilson

Status:	In Progress	**Start date:**	2014-05-29
Priority:	Normal	**Due date:**	
Assignee:	Bruce Wilson	**% Done:**	30%
Category:		**Estimated time:**	0.00 hour
Target version:			
Story Points:			

Description

Key questions will be what data would they expose through DataONE.

History

#1 - 2015-11-19 19:14 - Laura Moyers

- Status changed from New to In Progress

- Assignee set to Bruce Wilson

- % Done changed from 0 to 30

Last conversations with Line about PURR's involvement with DataONE were in June 2014.

Michael Witt was/is? the PURR PI and identified several pros and cons, which Line associated with the benefits to becoming a MN.

Michael's pros/cons:

Pros
1. NSF funding is expected to continue for another 5 years (not yet official) - too late to be named but could participate in second-hand funding if mutually compelling?
2. Participate in a community with a widely-recognized project
3. Can it be an opportunity to find researchers at Purdue who have data & want them included in DataONE, that this will motivate new collections in PURR? E.g., LARS? Maybe a local or smaller/targeted grant to make it happen?
4. Software development will be a good chunk of work but API is well-documented and uses some components that we already have in place
5. Will give us a third party who will replicate, ping, and verify integrity of our datasets
6. Datasets will be discoverable in ONEMercury catalog

Cons
1. Inadequate governance (bottom line: DataONE PI makes decision with input from advisors) - it is still an experiment, not yet a real organization or service
2. Not serious about preservation - not based on standards, no reciprocal agreements, would fail certification
3. Problems scaling (entire DataONE network is 90,000 files; we have a single dataset with 120,000 files in PURR)
4. Main benefits (increased access and high availability) require intermediating access through DataONE (e.g., ONEMercury Investigator Toolkit, or links to the DataONE website)
5. They have no cost model but need to be sustainable in 2019 at the latest, so they'll eventually have to hit us up for cash & we're already paying for MetaArchive & DPN
6. We don't have much relevant data in PURR for them right now (earth/life sciences) so they might not be interested in taking us (that said, I think they'd take us anyway to expand their base and our reputation)
7. Usage that comes through DataONE is logged by contributing nodes, not PURR (would need to download & re-integrate stats)

Line's translation:

Benefits of becoming a MN:
Reaching a wider audience: Pros 2.

Leveraging Existing Cyberinfrastructure through the Investigator's Toolkit: Michael did not mention anything related to this benefit. I do know that researchers have asked PURR to develop "handlers" to help select the data for download, i.e. Plug-ins that can preview the data before downloading, either for images or tables.

Is there anything like this planned with DataONE? How could a MN take advantage of this for their own data?

Receiving Recognition and Credit: PURR already achieves this goal by attributing DOIs

Maintaining High Availability: Pros 5

Enhancing collaborations: Pros 1.

And most importantly, how do I respond to the cons? 2 & 3 are pretty big. I feel that 2 is not quite fair, because the cyber-infrastructure needed to be put up, and DataONE decided to be agnostic w.r.t. Metadata standards to accommodate different Mns. So imposing a standard would not have been a good solution to attract Mns.

With 3, it would depend if an separate object is created for each file or if the dataset is regarded as a whole. Questions of mutability impact this.

#2 - 2015-11-19 19:17 - Laura Moyers

Bruce also had some feedback (June 2014) on Michael's pros/cons:

On the funding question (pro #1), the best route would be if there was an opportunity for PURR and DataONE to work on a (co-funded) project together. I'm not sure what that would be, but it bears some thinking about.

On the con side of things, particularly #1 and #2, I see where repositories, particularly ones like PURR, are coming from in terms of the organizational governance and the preservation standards. And becoming a more formal organization is on the roadmap for the next 5 year period. It's not quite "PI decides", in that we're covered by a cooperative agreement, not a grant. But I don't think that qualitatively addresses Mike's concern on that point. In the end, the data resides with the member nodes, and they're largely the ones most directly responsible for the preservation of the data. We have a broad range of groups participating and working to participate, some of whom are very much organizations that can pass a preservation audit and others that are much less mature as repositories. I view DataONE as something that helps facilitate preservation. However, given the feedback we got from several people (not just from Mike) on this point, it's something I'll be bringing up with project leadership.

On con #3, I need to check some stats. However, there are some places where we've got to address scaling, and an XML file with 120,000 nodes is problematic. However, a dataset that has 120,000 nodes is problematic, at least from a user interface perspective, no matter what the system. It's not clear to me that there's any standard pattern for dealing with data sets of that high degree of granularity. Having said that, however, we have run tests of Mercury against systems with 100 million nodes and it scales just fine in terms of performance. That's only part of the answer, however, as the rest of the architecture has to scale. We can scale out, in that each CN site can actually have multiple VM's behind the scenes. It should scale, but I stipulate that's an assertion, not a statement backed by data and large scale testing.

On con #7, data download stats only come through DataONE where there's replicated content. Anything that's in PURR and is accessed directly at PURR is still part of your logs.