OLRC Update

2016 OCUL Digital Curation Summit, Carleton University

Cameron Metcalf (University of Ottawa)

October 27, 2016.

Overview

- What is the OLRC? (~4 slides)
- Recent survey response from participating institutions (~16 slides)
- Future opportunities (~1 slide)

What is the OLRC?

- **Ontario Library Research Cloud**

- **A Cloud service:** the Ontario Library Research Cloud (OLRC) provides preservation storage for Ontario's scholarly material on five regionally distributed nodes.

- Launched October 2015

- Contact: cloud@scholarsportal.info

Service Objectives*

- Provide cost-effective subscription-based scalable storage for Ontario's academic libraries to house valuable and expanding digital collections
- Provide preservation services for that content to ensure long-term access
- Develop new tools for researchers to be able to explore and analyze that content at scale

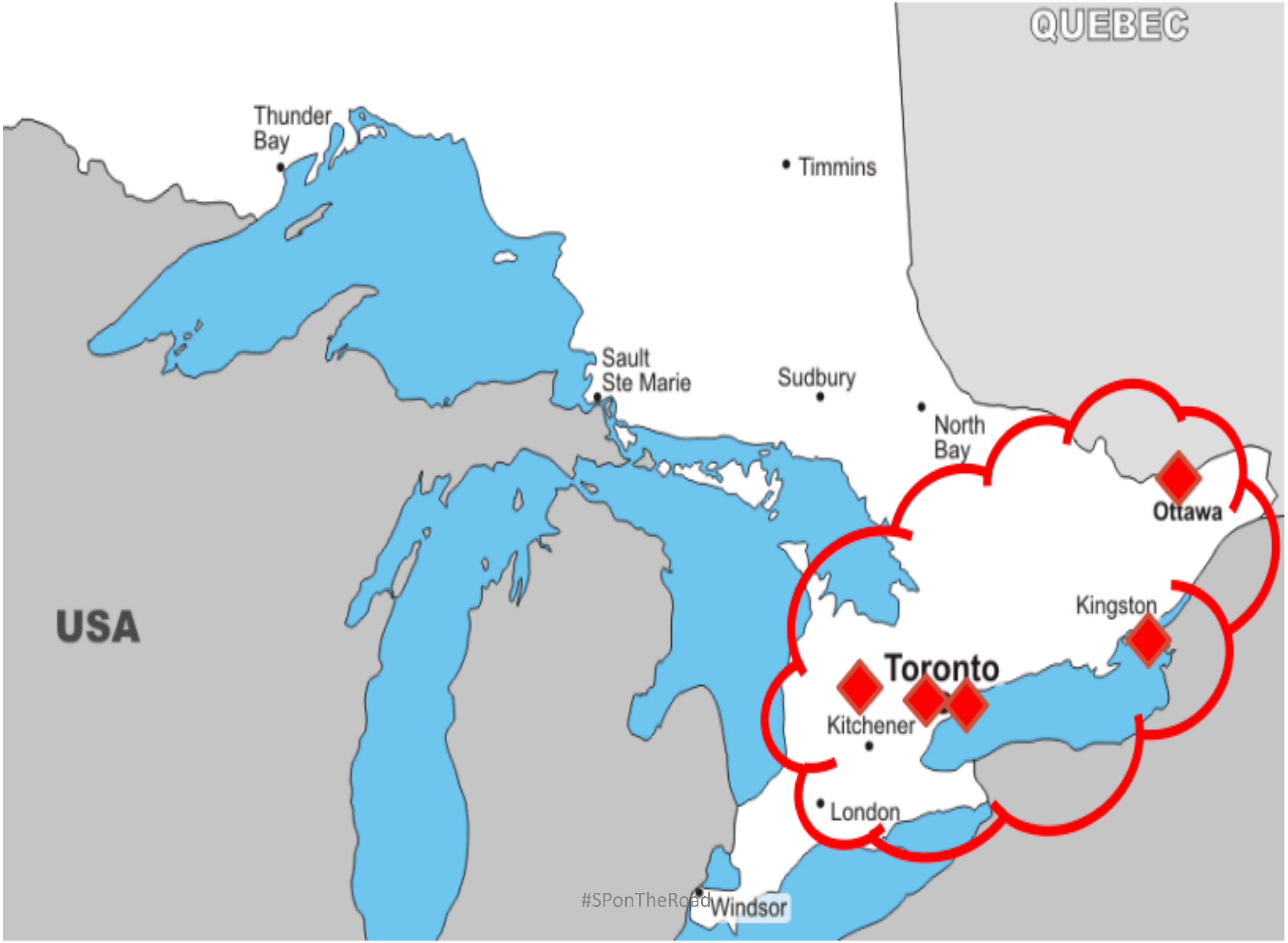
* From the Scholars Portal Roadshow 2016

The Cloud *

- Utilizes high-speed virtual network running on ORION and GTAnet to provide 10G connectivity between 5 hosting sites: *Toronto, Ottawa, Queens, York, & Guelph*
- Digital objects are replicated in the cloud at least three times to ensure redundancy (no single point of failure)
- Powered by OpenStack Object Storage (Swift) software

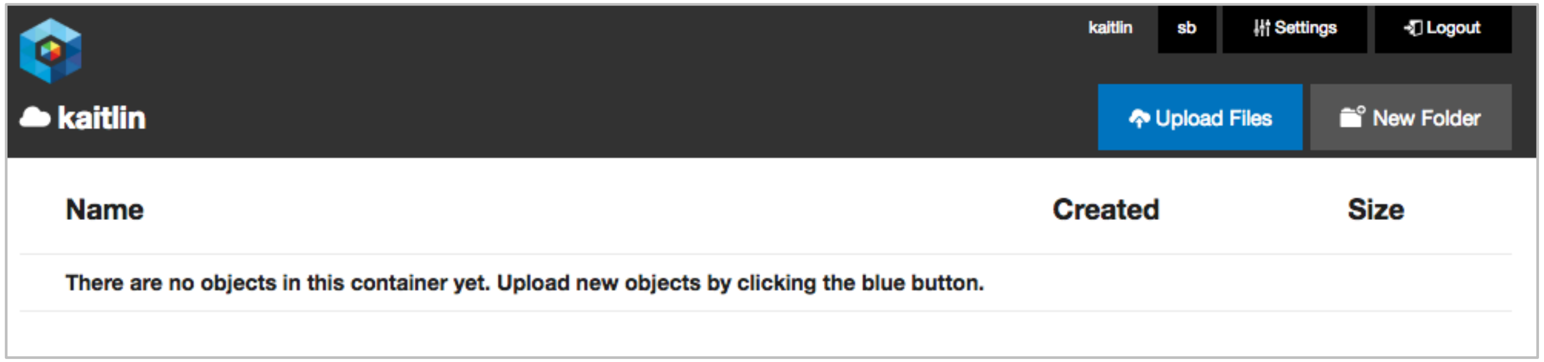
*From the Scholars Portal Roadshow 2016

openstack™
CLOUD SOFTWARE

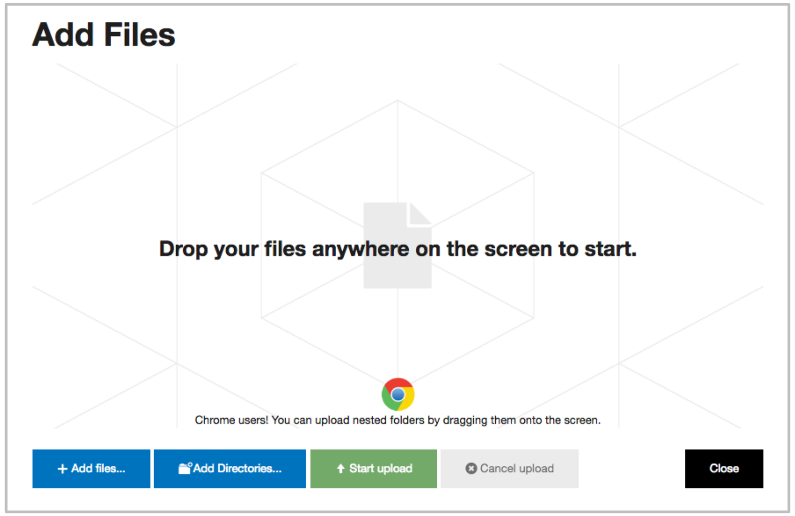


#SPonTheRoad

Swiftbrowser

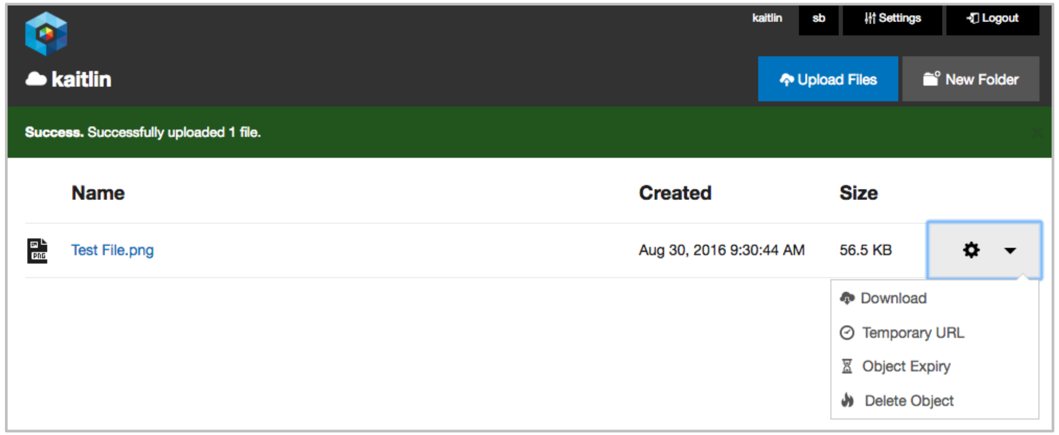


Name	Created	Size
There are no objects in this container yet. Upload new objects by clicking the blue button.		


Add Files

Drop your files anywhere on the screen to start.

Chrome users! You can upload nested folders by dragging them onto the screen.

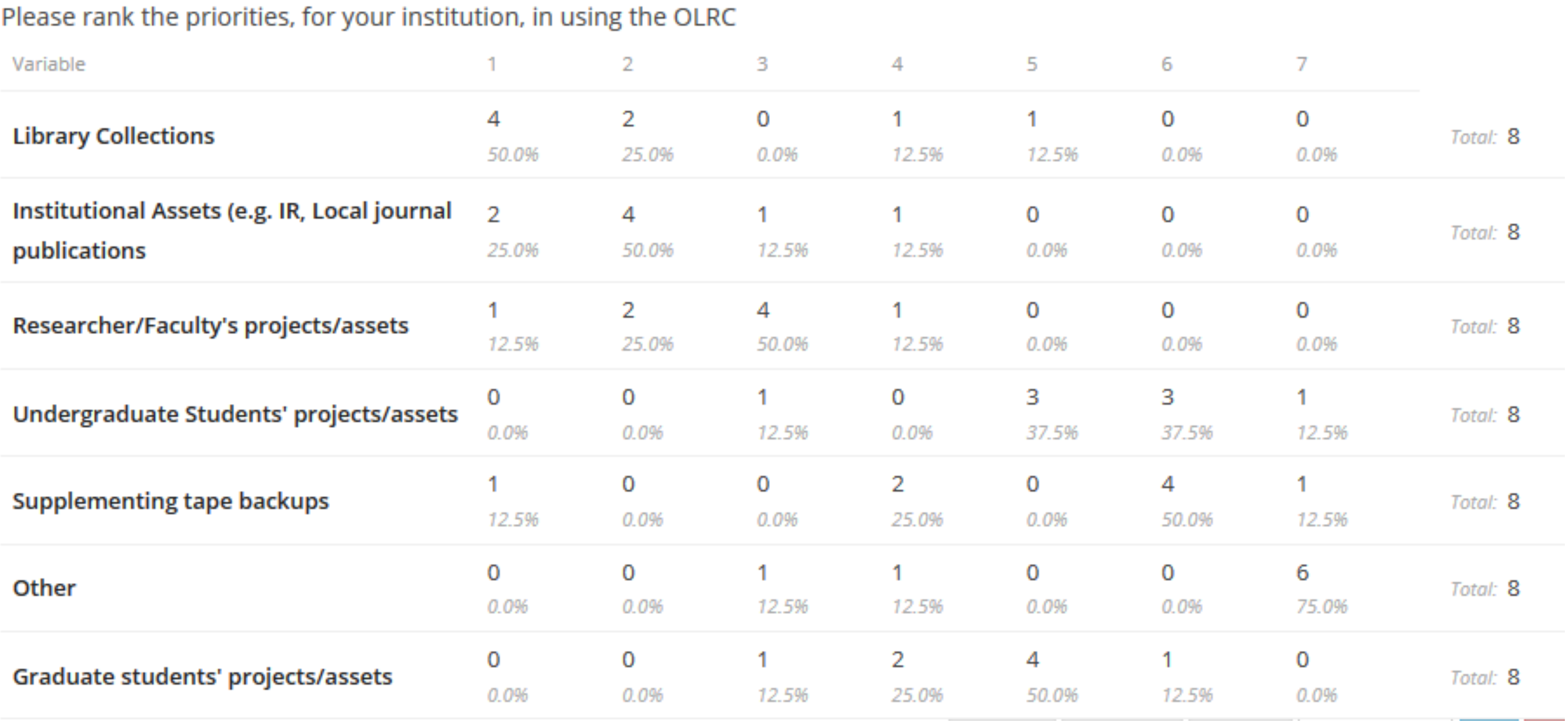


Success. Successfully uploaded 1 file.

Name	Created	Size
Test File.png	Aug 30, 2016 9:30:44 AM	56.5 KB

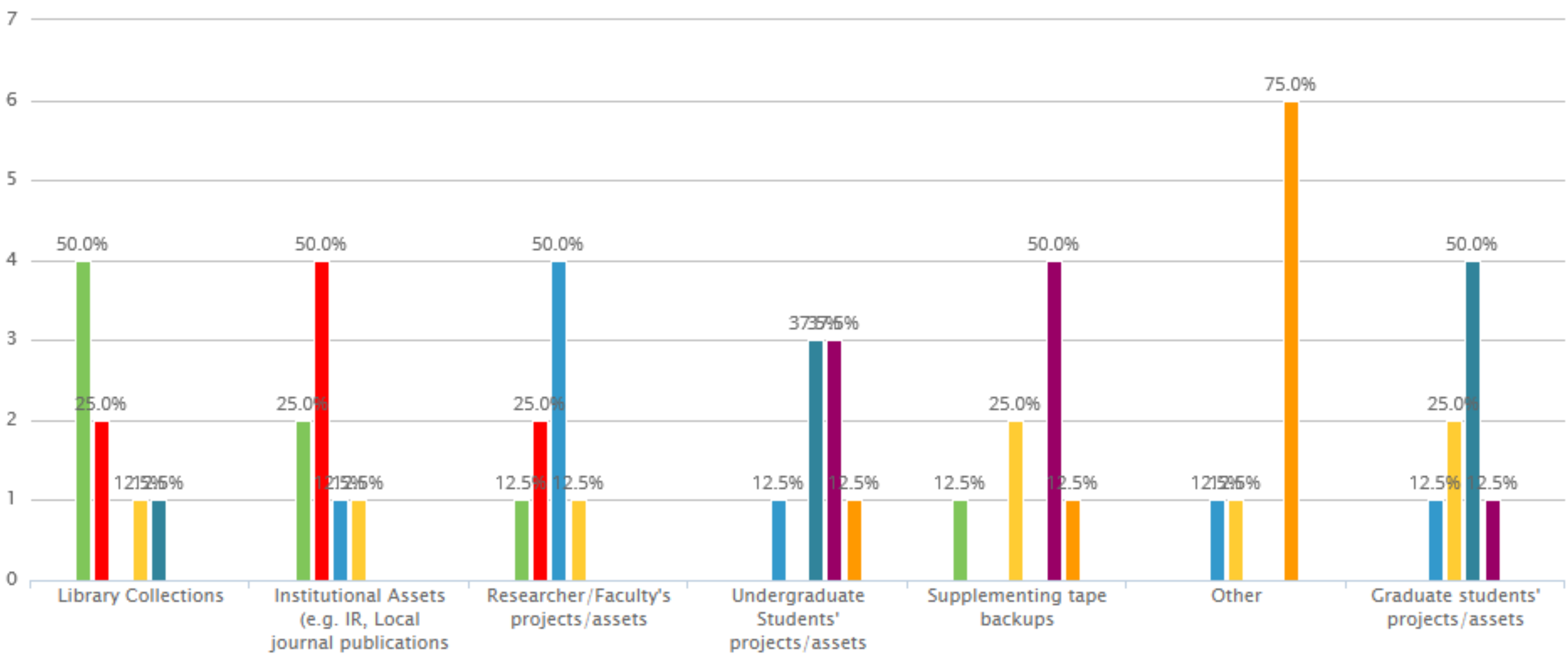
OLRC Survey - from this past week

- Carleton University
- McMaster University
- Queen’s University
- University of Guelph
- University of Ottawa
- University of Waterloo
- University of Windsor
- Wilfrid Laurier University



Please rank the priorities, for your institution, in using the OLRC

Variable	1	2	3	4	5	6	7	
Library Collections	4 *50.0%*	2 *25.0%*	0 *0.0%*	1 *12.5%*	1 *12.5%*	0 *0.0%*	0 *0.0%*	*Total:* 8
Institutional Assets (e.g. IR, Local journal publications	2 *25.0%*	4 *50.0%*	1 *12.5%*	1 *12.5%*	0 *0.0%*	0 *0.0%*	0 *0.0%*	*Total:* 8
Researcher/Faculty's projects/assets	1 *12.5%*	2 *25.0%*	4 *50.0%*	1 *12.5%*	0 *0.0%*	0 *0.0%*	0 *0.0%*	*Total:* 8
Undergraduate Students' projects/assets	0 *0.0%*	0 *0.0%*	1 *12.5%*	0 *0.0%*	3 *37.5%*	3 *37.5%*	1 *12.5%*	*Total:* 8
Supplementing tape backups	1 *12.5%*	0 *0.0%*	0 *0.0%*	2 *25.0%*	0 *0.0%*	4 *50.0%*	1 *12.5%*	*Total:* 8
Other	0 *0.0%*	0 *0.0%*	1 *12.5%*	1 *12.5%*	0 *0.0%*	0 *0.0%*	6 *75.0%*	*Total:* 8
Graduate students' projects/assets	0 *0.0%*	0 *0.0%*	1 *12.5%*	2 *25.0%*	4 *50.0%*	1 *12.5%*	0 *0.0%*	*Total:* 8



- **Green** = 1st choice
- Red = 2nd choice
- Blue = 3rd choice

- **Other: 6 institutions ranked it as 7th priority**

Comment to define: "unprocessed, i.e.- non-accessioned, born-digital in-kind donations."

The type of assets being stored

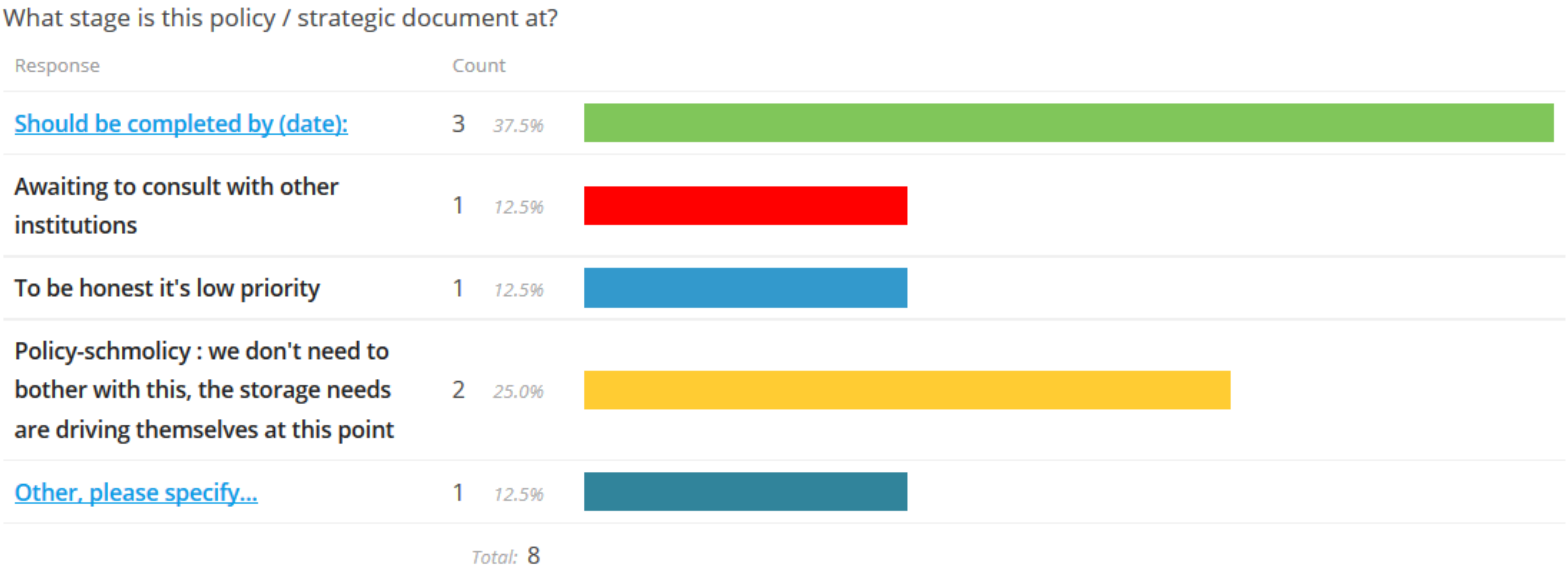
- digitized images, audio recordings, research data, archived web content
- maps, scanned manuscripts, scanned books, IR content, journals, misc. digital files donated to us
- images, video, geospatial data, digital objects from the library archives, Library staff intranet back-ups
- Most of our uploads have been scanned images.
- Scanned images from archives, videos from special collections, scanned documents, ETDs, etc.
- PDFs, oral histories, databases, image files
- digitized collections, ETDs, faculty digital proejcts
- scanned images, research data, internal collections

Has your institution worked on an internal policy/strategic document outlining your approach to using the OLRC?

(i.e. what you would include / how you would prioritize what is uploaded)

Yes = 6

No = 2

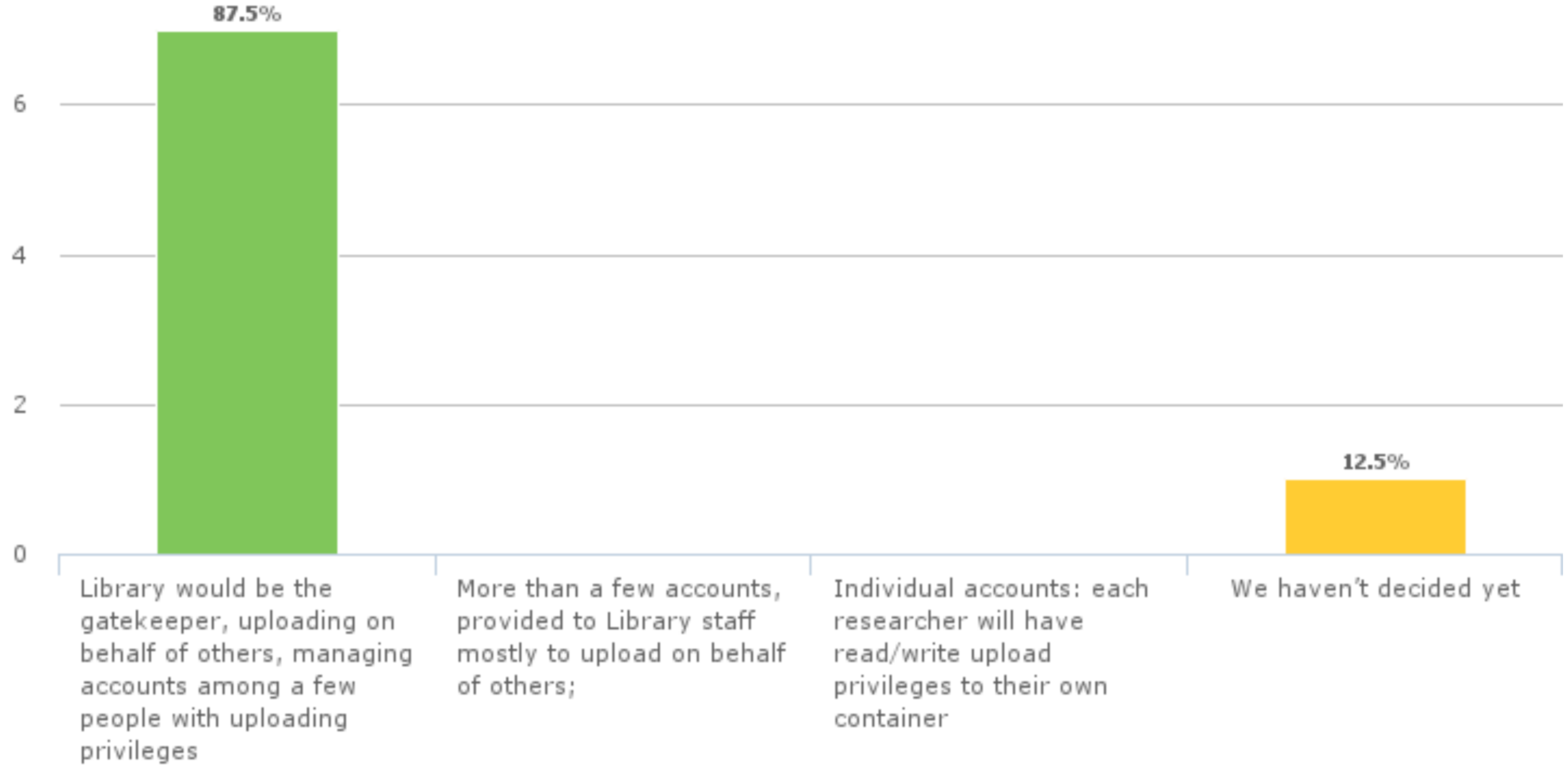


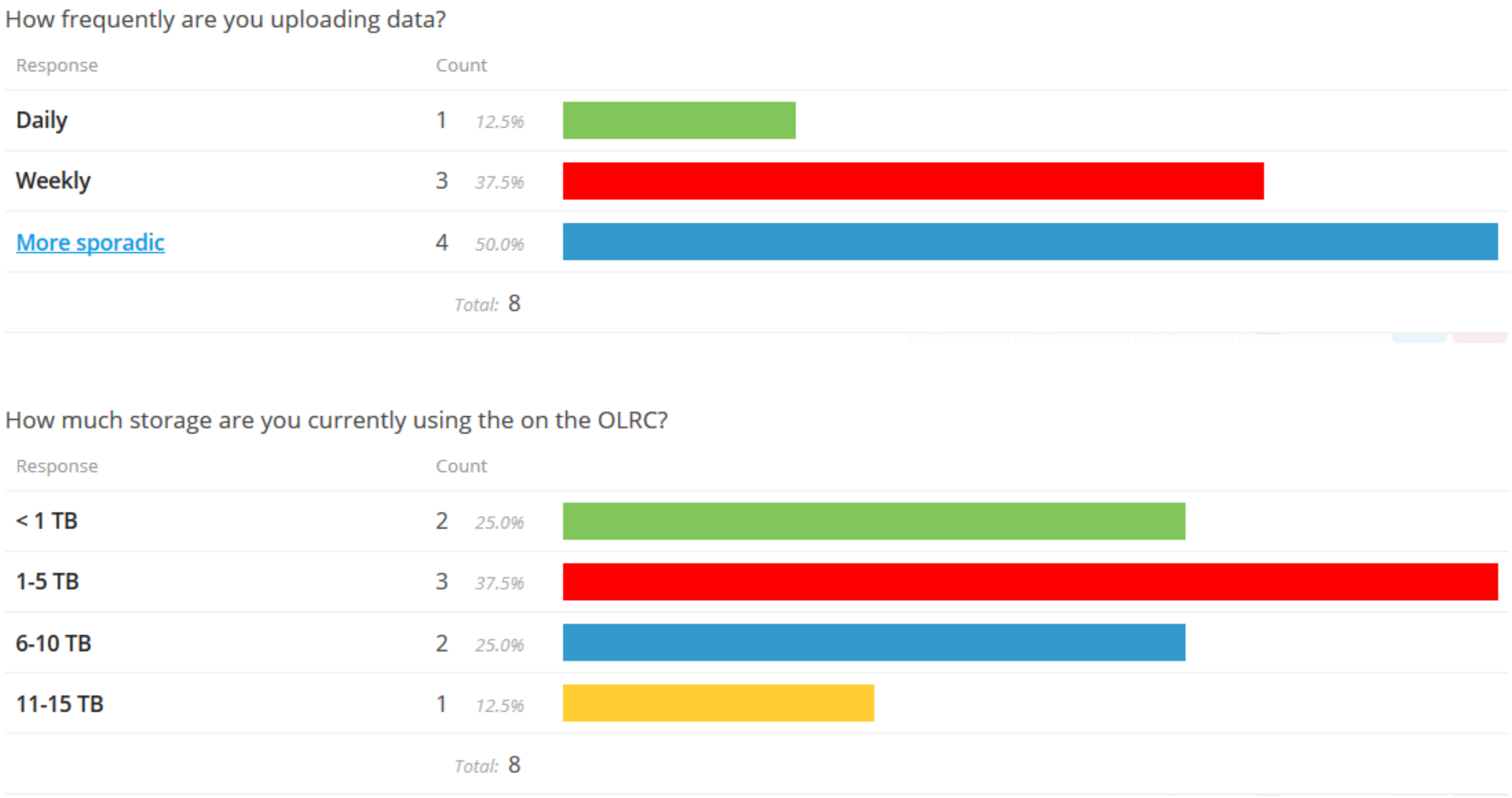
What stage is this policy / strategic document at?

Response	Count	
Should be completed by (date):	3	37.5%
Awaiting to consult with other institutions	1	12.5%
To be honest it's low priority	1	12.5%
Policy-schmolicy : we don't need to bother with this, the storage needs are driving themselves at this point	2	25.0%
Other, please specify...	1	12.5%
	Total: 8	

- completion date = December 2016, for all three.
- rough Notes from meetings, to start drafting a policy.

What is your Library's approach to providing access, so far?





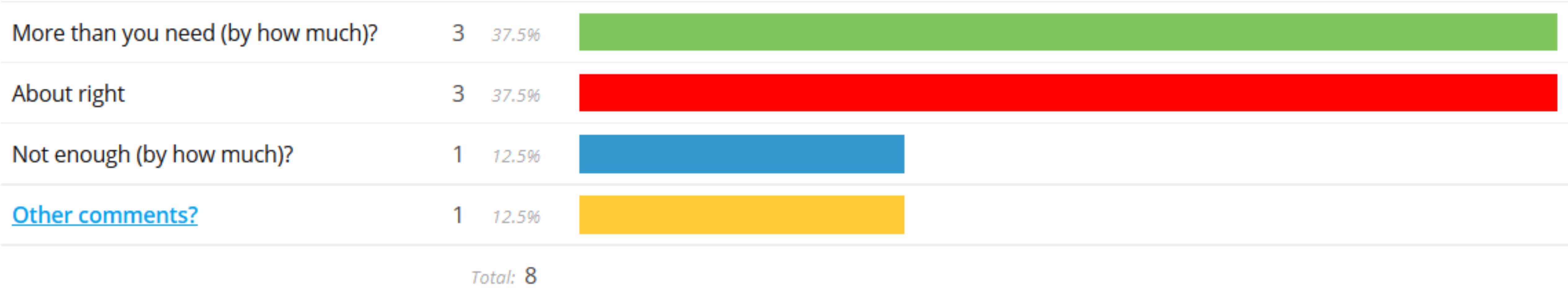
How frequently are you uploading data?

Response	Count	
Daily	1	*12.5%*
Weekly	3	*37.5%*
[More sporadic]	4	*50.0%*
	Total: 8	

How much storage are you currently using the on the OLRC?

Response	Count	
< 1 TB	2	*25.0%*
1-5 TB	3	*37.5%*
6-10 TB	2	*25.0%*
11-15 TB	1	*12.5%*
	Total: 8	

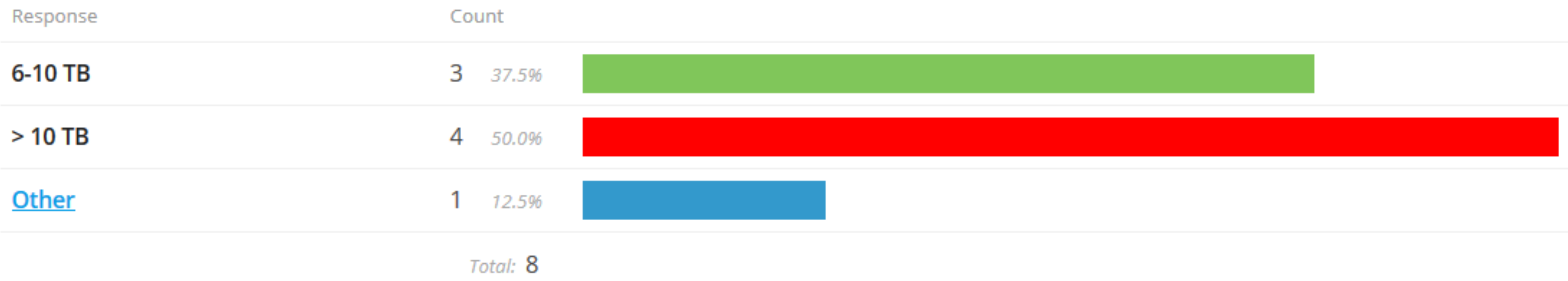
Everyone has 20 TB right now. Is it …



More than you need (by how much)?	3	37.5%
About right	3	37.5%
Not enough (by how much)?	1	12.5%
Other comments?	1	12.5%

Total: 8

How much do you expect to use 1.5 years from now?



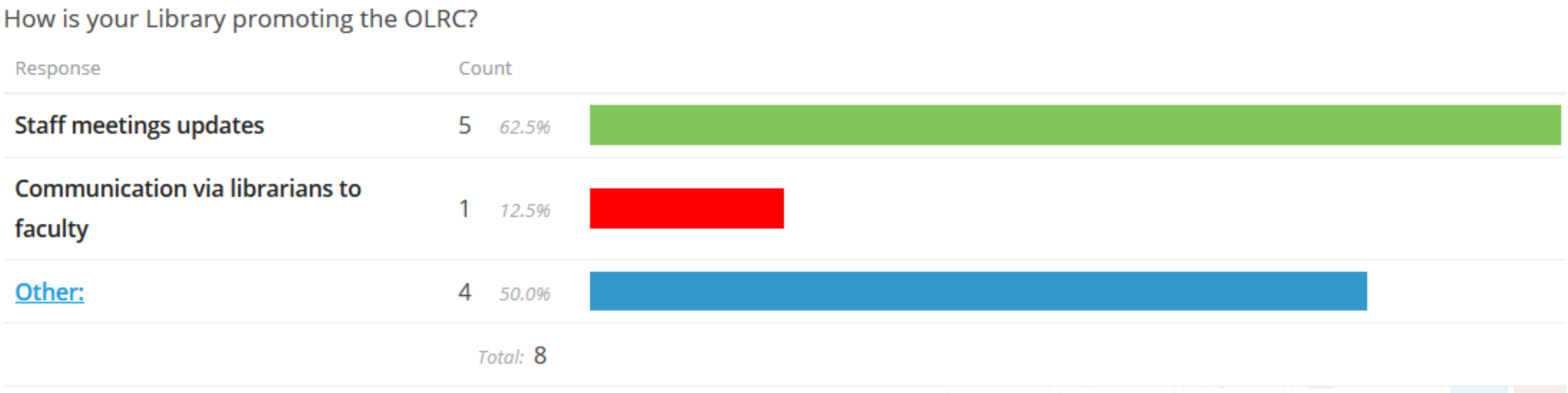
Response	Count	
6-10 TB	3	*37.5%*
> 10 TB	4	*50.0%*
Other	1	*12.5%*
	Total: 8	

Only one out of Eight

1 of 8 institutions reported that they’d encountered significant technical barriers due to dropped connections for larger files.

Their solution:

They modified the file uploader to pass on authentication credentials more frequently and prevent any lapse in recognition during these extended uploading sessions.

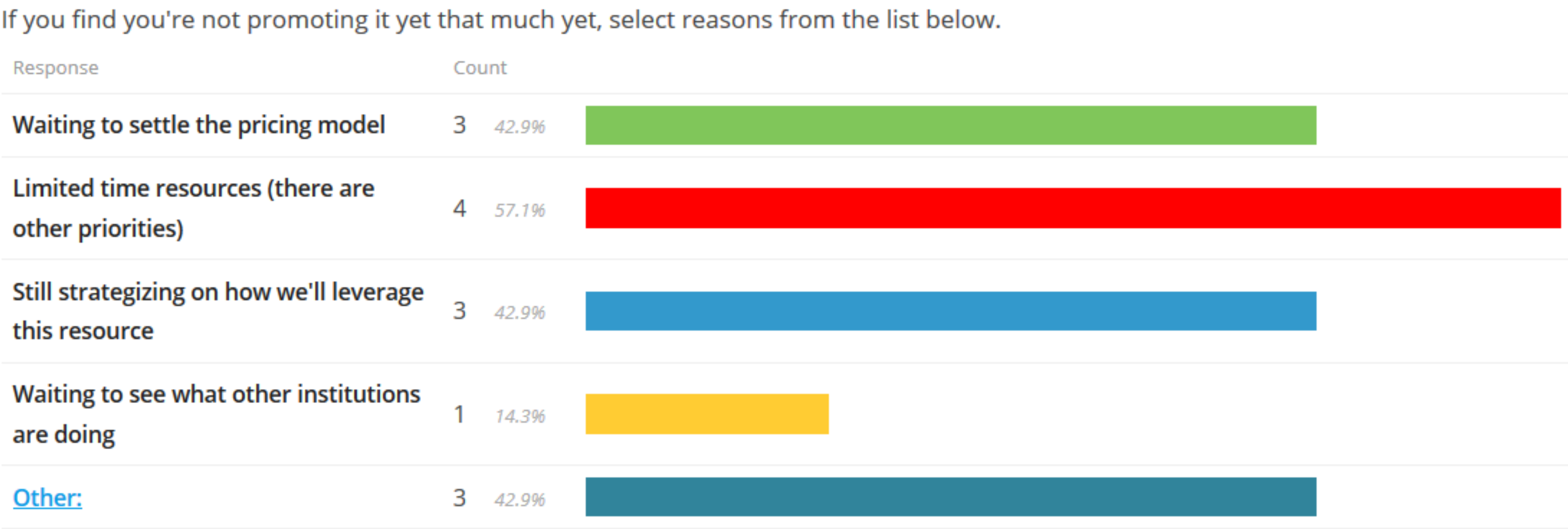


How is your Library promoting the OLRC?

Response	Count	
Staff meetings updates	5	62.5%
Communication via librarians to faculty	1	12.5%
Other:	4	50.0%
	Total: 8	

* Other:

- Meetings with key stakeholders
- Not really promoting yet, a mention here & there
- conference presentations
- nothing at this time



If you find you're not promoting it yet that much yet, select reasons from the list below.

Response	Count	
Waiting to settle the pricing model	3	42.9%
Limited time resources (there are other priorities)	4	57.1%
Still strategizing on how we'll leverage this resource	3	42.9%
Waiting to see what other institutions are doing	1	14.3%
Other:	3	42.9%

- other:

  “Waiting until the policy is formally endorsed”

  “We are using it strictly for our preservation program at present. That's an internal concern although it helps us fulfill the promise we make with our IR and OJS platforms that we will preserve the data well into the future.”

  “Don’t have a service extending beyond the library to promote yet”

What is the current primary benefit of the OLRC?

- A research project has been conducting audio interviews with Native women in remote areas. As these interviews will inform the multi-year project it is critical that they have robust storage. And since there is a high need for confidentiality the storage must also be highly secure. The PI has much greater piece of mind knowing the recordings have been archived in the OLRC.

- We now have a viable, replicated, offsite storage solution for our unique digital assets, in a place where we have relative cost security and sanity. Also, it's in Canada so we bypass any of the unpleasant "where's your data" conversations.

- We have been able to back-up and store collections that had previously been sprawled around various hard drives and CD/DVD storage

- ODW (OurDigitalWorld) newspaper project would be almost inconceivable at this point without OLRC.

- Ability to back-up our growing GIS data; ability to start planning for preservation of digital repositories

- Encrypted backups of many of our important systems is very nice. A few one off dark storage / preservation projects are much easier to complete.

- Saying we have the OLRC at our disposal

What future enhancements do you await?

- Archivematica and Dataverse integration; enhanced capabilities in the uploading options
- Perhaps some sophisticated file integrity checking tools that make fixity checking, etc., much more of an automated and transparent process.
- Options for Archivematica and OwnCloud are intriguing for us. On our side, we want to establish process and policies for biling and mounting data to make it a quicker/easier procedure to provide this as a real viable service to researchers looking for data space at a manageable rate of growth.
- By far the biggest enhancement(s) would be a CAS authentication mechanism and dropbox-like integration with the desktop. Dropbox is pervasive on campus among all groups, including Faculty, for sharing data and backing up content. OLRC could be a much better option but ease of use is critical to compete. Userid and passwords are also terribly hard to manage for the library, it makes much more sense to fall back on existing authentication which is also consistent with other campus and inter-campus services, e.g., eduroam.
- Archivematica integration, Dataverse integration, Data mining investigations with Scholars Portal
- Archivematica integration. Individual accounts for users, with limited sized buckets.
- Archivematica to normalize workflow for using it
- Archivematica, DataVerse, Islandora

Two out of Eight

2/8 institutions worked on customizations with the OLRC interface options:

- We had done quite a bit of work with the openstack compression option, a method of using widely implemented tar gzip tools for more efficiently moving data across the network. We had originally explored storing derivative files from the newspaper project on OLRC, including PDFs and tiles, but decided against this approach since: Strictly speaking, the most critical object is the source file and in a worst case scenario, the derivatives could be regenerated from scratch in the event of data loss Windsor stores over 70 TB of newspaper data for the ODW newspaper initiative and this collection if constantly growing. Storing everything associated with the newspaper project in the OLRC would require a lot of disk space, far beyond the 20 TB threshold. Even with files in tar.gz format, transferring files is very time-consuming, it would have taken many months to get everything into OLRC.

- Small patch to large file uploader tool.

Last words: suggestions for improvement

- Very much in line with this survey, I think it would be highly beneficial to know what other institutions are doing with the OLRC. Perhaps some communities of interest could be formed. I'm particularly intrigued by the notion of local enhancements (e.g. using the API). I also find a certain amount of ambiguity around the OLRC as a dark archive versus the OLRC as a space where research can be conducted. (e.g. analysis, mining, visualization) Some common understanding or recognition of different perspectives would be helpful.

- Not really; we just hope to see everyone in OCUL using it for preservation purposes sooner rather than later. We are willing to help by sharing our process and procedures.

- Ongoing information-sharing among institutions to see what others are doing to leverage those ideas for ourselves.

- An alternative to dropbox with institutional authentication would be highly welcome.

- More clarity around status, how others are using it, suggestions on workflows, suggestions on best way to start

- Vanity urls for public resources that don't have AUTH_ONEMILLIONCHARS in the URL would be nice.

Future Opportunities

Cost per TB as of May 2017*

- Annual subscription: $320/TB for OCUL members, $790/TB for non-OCUL
- OCUL schools pay a $1600/yr 5TB minimum,   Non-OCUL clients pay a $3950/yr 5TB minimum
- 3% discount for 10TB block
- 6% discount for 50TB block
- **Free 5TB Trial until April 2017**
- Contact: cloud@scholarsportal.info

*From Scholars Portal Roadshow 2016