DuraCloud DSpace Interaction Notes

Most of these comments originate from discussion between Richard Rodgers, Bill Hays and Tim Donohue on 15 April 2010. Feel free to enhance or add your own notes.

DuraCloud Synchronization

How does the "DuraCloud Sync Tool" (which watches a file system folder and synchronizes with the cloud) actually work?

In current scenario, you need to export all of DSpace into AIPs in a local file system folder, and tell the DuraCloud Sync Tool to synchronize that folder into the cloud. This is inefficient as it requires you to replicate all your content locally (to the sync folder) **before** it can be replicated to the cloud via DuraCloud. In other words, you'd now have 3 copies of this content: (1) in DSpace, (2) as exported AIPs in your local sync folder, and (3) as exported AIPs in DuraCloud.

1. Does DuraCloud automatically synchronize all changes in the local folder? For instance, if a file is deleted, is it removed from the cloud storage?
   - **Tim's Answer:** I talked to Bill Branan from DuraCloud team. The current implementation of the Sync Tool always synchronizes with local folder contents. So, if you delete a file from that folder, it will be removed from the cloud storage. However, after our discussions of use cases, Bill agreed it may be necessary to have a way to "turn off" the auto-delete functionality. So, that if you remove a file locally, it will **n ot** auto-delete it from the cloud (unless you explicitly force the delete).
2. Would it be possible to perform a "trickle" synchronization for large amounts of content? For example, if your DSpace has 1TB of content, you wouldn't want to export the entire 1TB at once locally (thus doubling your local storage needs). Rather, maybe it would be possible to export 10GB at a time to a local DuraCloud Sync Folder, and have that content "trickle" up into the cloud.
   - **Tim's Answer:** Again, based on discussion with Bill Branan. Currently, the DuraCloud Sync Tool doesn't support this sort of "trickle" synchronization. However, it could support it if there was a way to turn off "auto-delete" in the Sync Tool (so that it would no longer auto-delete content in cloud which has been removed from the local sync folder). *See below for more details*

Possibility: Adding an option to turn off "auto-delete" in DuraCloud Sync Tool?

In the end, it sounds like our best option may be to ask the DuraCloud Team to create a way to turn off the "auto-delete" option of the DuraCloud Sync Tool. Bill Branan seemed open to this idea, and is investigating it (from the sounds of it, he thought it'd be a worthwhile and hopefully not too difficult of a change).

Here's a basic example future DSpace/DuraCloud interaction workflow, that may allow us to perform a "trickle" synchronization of content:

1. Create a Folder, and turn on DuraCloud Sync Tool and point at that Folder. Make sure the "auto-delete" option of the Sync Tool is turned OFF.
2. Export the first 10GB of AIP content from DSpace into that Folder. DuraCloud will automatically notice the new files and sync them up into the cloud storage.
3. Next, remove the already synced 10GB of content from the Folder (if "auto-delete" option is turned OFF, DuraCloud will retain that content in cloud storage).
4. Export the next 10GB of AIP content from DSpace into that Folder. DuraCloud will automatically notice the new files and sync them up into the cloud storage. This will bring the total storage in DuraCloud up to 20GB (even though your local sync folder only has 10GB)
5. Repeat in batches of 10GB until all of DSpace content AIPs are loaded into DuraCloud

Obviously, there are a few things to note about this workflow:

1. If "auto-delete" was turned OFF, DuraCloud will **never** remove any content, until you explicitly tell it to. This means we may need a "cleanup" script or have an "audit" script which can clean up unnecessary files that still exist in DuraCloud that were removed from DSpace.
2. DuraCloud will accept updates to files. If you place a file into your sync folder with the name "ITEM-123454678-1.zip", and DuraCloud already has a file of that name in its storage space, DuraCloud will compare the files (via checksum). If they are different, the new file will overwrite the old file.
3. **BIG ISSUE** If you accidentally switched the "auto-delete" option back ON, DuraCloud may auto-delete all/most of your content. Bill Branan & I discussed this as a major concern that we need to resolve in some way. Maybe DuraCloud Sync Tool needs to default to **not** auto-delete content? Or, at the very least, explicitly WARN you if you tried to turn ON "auto-delete".

Auditing Functionality

Once content is in DuraCloud, we need a way to audit that content and compare it to what is currently in DSpace.

A very simple DuraCloud/DSpace auditing workflow may be as follows:

1. Export an AIP for a random DSpace Object (or a chosen one) to local filesystem
2. Generate a local checksum of the exported AIP
3. Using the DuraCloud REST API, compare that local checksum with the checksum for that item as stored in DuraCloud
   - If the checksums match, then the content is identical (successful audit)
   - If they don't match, then you know one or the other is out of sync

4. Repeat as necessary for some/all objects in DSpace