DuraCloud Guide

This page provides an overview of DuraCloud and answers common questions about the service. You can also review full documentation for DuraCloud or learn more about subscriptions from DuraCloud by LYRASIS or DuraCloud Europe from 4Science. You can download a flyer about our services (PDF).

What is DuraCloud?

DuraCloud is an open source, hosted service that makes it easy to control where and how your organization preserves content in the cloud. DuraCloud enables your institution to store content with expert cloud storage providers while adding lightweight features that enable digital preservation, data access, and data sharing. The service is available from LYRASIS or from 4Science via DuraCloud Europe.

DuraCloud is designed to meet the needs of cultural heritage institutions, with features including:

- Storage and replication of content across multiple providers via a desktop tool, REST API, command line, or through a web-accessible interface
- Bit integrity health reports on all content at least twice per year
- Multiple content transfer interfaces, with options for both novice and technical staff
- Predictable annual billing and the option to add additional storage at any time
- Flexibility to combine private storage, public access, and dark archive options
- Integrations to support repository backup, archival file storage, and website archiving
- Optional administrative access controls

Content Organization, Access, and Metadata in DuraCloud

How Content is Organized

Within DuraCloud, content is organized into containers called spaces. Each institutional account can include up to 100 spaces, with the option to increase this number if the need should arise. Access controls are at the space level, so Enterprise-level account administrators can create user groups and control who has read or write access to a given space. Content is transferred from the local system to a specific DuraCloud space using one of the transfer methods described below.

When a directory of files (which can include sub-folders) is uploaded to DuraCloud, the original file structure is maintained in the name of the item. This means that when a directory is later retrieved from cloud storage, the original structure (i.e. folders) is replicated. This also makes it easier to locate specific items based in the file name, as DuraCloud does not currently allow for searching within a space. A search feature is frequently requested and is a high priority for future development.

Public Access Option

Spaces within DuraCloud are private by default, with files only accessible to authenticated and authorized users. An institution can choose to make a space public. Every item in a public space has a URL. This URL can be used to grant access to the item through a repository, CMS, or website.

File Types, Metadata, & Directory Configuration

There are no requirements on how your content must be structured for ingest into DuraCloud. DuraCloud is capable of storing any type of file or package (i.e., AIP, ZIP, TAR, etc.).

DuraCloud does not require any specific metadata schema. Through the DuraCloud web interface or REST API, you can add as many different name /value pairs of metadata as you need, on a content item or DuraCloud space basis. You can also tag your content stored in DuraCloud in the same way.

Metadata Captured During Transfer

As the Sync Tool transfers files to DuraCloud, it will attempt to capture certain types of metadata about each file, and include that information as part of the content item added to DuraCloud. See this list for a full description of the metadata that is captured automatically. You have the option to add, update, or delete the properties of each file after it has been transferred to DuraCloud.

DuraCloud Best Practice: Configuring Your Directories for Syncing to DuraCloud

Deciding how to sync directories to DuraCloud is an important step. You may wish to have a directory structure with a minimum depth of two levels inside the synced folder since it will allow you to see a deeper and more meaningful directory hierarchy in the DuraCloud dashboard. To retain hierarchical information in your filenames, you can either create at multiple folder depths for syncing or use the DuraCloud prefix option in the SyncTool settings. If the order and intellectual context of your content and/or its placement in your original filesystem is essential to its authenticity over time, these instructions will help reflect that.

Create a new directory to sync:

1. Create a top level directory on your local machine or server to sync to DuraCloud. In the example below, it's called 'syncFolder'. The name of this top level folder will not be included in the filename once you've uploaded into DuraCloud.
2. Be sure to create all of the meaningful levels of directories within that top level directory before any digital objects.
3. In this example, the directory called OralHistories, the directories below it, and the filename will be visible in the DuraCloud dashboard. In DuraCloud, your file would appear as OralHistories/Person1/digital.wav

```
/syncFolder
   /OralHistories
      /Person1
          digital.wav
      /Person2
          thing.wav
```

Sync an existing directory:

1. If there are existing directories that you wish to sync but do not wish to reorganize them into or add them from a directory structure deep enough to reflect the hierarchy of your content, you can instead use the Sync Tool's prefix option.
    a. Ensure your sync tool is stopped in the Status tab. Then click on the configuration tab.
    b. Under "Other options," create a prefix for your sync folder to create a directory structure at least two levels deep. The prefix must end in a slash (/)
2. Note that the prefix will replace the directory name of the sync folder in DuraCloud. If the sync directory selected in the sync tool is `syncFolder`, you could add a prefix such as: OralHistories/Person1. In DuraCloud, your file would appear as OralHistories/Person1/DigitalStuff/digital.wav

```
/syncFolder
   /DigitalStuff
 /digital.wav
```

Content Transfer to & Retrieval from Cloud Storage

Web Interface

Users can interact with a browser-based graphical user interface to view and manage content in DuraCloud. The web interface offers access to all storage system capabilities, including space and content creation, updates, and deletion. It provides access to graphical depictions of the information contained in the storage reports and allows for bulk deletion of spaces and content items and for user account administration. Administrators can use the interface to designate read and/or write access to a given space for a given user or group of users.

DuraCloud SyncTool

The SyncTool provides a simple way to move files from a local file system to DuraCloud. The Sync Tool provides a web-browser-based application user interface which begins with a configuration wizard, then provides a dashboard display showing the current status of the sync process. This interface is the default and is started by selecting any of the shortcuts created by the installer. The user can select directories and sub-directories for the SyncTool to either upload from once in a single pass, or can set the tool to watch and sync the directories to the space automatically.

Please review the Sync Tool minimum requirements and recommended best practices for configuring your directories for the SyncTool.

Command Line

A command line interface for the SyncTool is also available. It can be executed directly, used in scripts, or used for scheduling sync activities (e.g. within a Cron Job.) The command line interface provides access to all features of the SyncTool, some of which are not currently available in the graphical interface.

Chunking Files

The DuraCloud SyncTool will "chunk" files as they are sent to DuraCloud. What this means is that if a file is over a pre-defined size limit (by default this is 1GB, but can be set up to 5GB in the tool configuration settings), that file is transferred in segments. A checksum for each segment is generated and captured in a manifest for the file which also includes the checksum for the entire file. When the entire file has been transferred to DuraCloud, you will see the list of chunks as well as the manifest file in storage.

Chunking and Stitching in Detail

To chunk a file the SyncTool reads bytes in the source file and writes them into a temp file on the local file system until reaching the defined chunk size limit (by default, 1GB.) At that point the temp file becomes the first chunk, so the SyncTool will compute the checksum and transfer the file to storage. The process then continues to read bytes from the source file into a new temp file, compute its checksum and transfer it to storage, and the process repeats until reaching the end of the source file. Along the way, as each chunk is created the SyncTool will write the details for that chunk into a file DuraCloud calls the "chunks manifest"; this includes the name of each chunk file (which is the original file name plus a numbered suffix) and its checksum. When all chunks have been transferred the chunk manifest file is finalized and is transferred into storage.

Stitching files is essentially the reverse of chunking. The Retrieval Tool will first pull and read the chunk manifest file to determine all the chunk files that are needed to construct the original file. It will then pull the first chunk and write it to disk, then the second file, which is appended to the first, then the third, and so on until each of the chunks have been appended to the end of the file. As each chunk is pulled the system will check its checksum to verify it was downloaded correctly before appending it. Given that it is working at the byte level, this is simply constructing the same stream of bytes as the original file. A final checksum comparison with what is recorded in the chunks manifest is used to verify that the completed file is consistent with the original checksum.

REST API

Authorized users can also choose to interact with a DuraCloud account through the REST API. The API offers the same functionality as the graphical user interface. Complete REST API documentation is provided.

Retrieval Tool

Regardless of which provider you choose, you can at any time use the Retrieval Tool to download your content is bulk from DuraCloud. This tool ensures that the content you download from DuraCloud looks the same as what you added. That includes recreating the original directory structure, stitching all chunked files, and even re-setting the timestamps on those files to the original values (where possible.) If retrieving content from Chronopolis dark archives storage, there is a per-TB fee for retrieval.

Chronopolis Network Deposit Option

Chronopolis Deposit Process

Content is uploaded into the DuraCloud system via the DuraCloud SyncTool and is initially stored in Amazon Web Services S3 storage. This allows for initial checksums to be performed and for the client to stage and organize content prior to ingest into the Chronopolis system. An authorized administrative user initiates the ingest process, called the “snapshot.” clients are asked to move content from DuraCloud into the Chronopolis System within 3 months of upload or to pay the prorated cost $700/TB/year for the additional time client Content is stored in Amazon S3.

After the client initiates the Content and Data ingest process (i.e. the “snapshot”), client’s data will be stored in a digital preservation system of geographically distributed nodes, each managed using vendor supported software and hardware.The architecture of this system allows for the failure of an entire Chronopolis node with data still available and reliable at the other nodes.

Chronopolis Data Validation

All data in the system will be monitored for fixity using the Auditing Control Environment software (ACE). ACE is a system that incorporates a new methodology to address the integrity of long term archives using rigorous cryptographic techniques. ACE continuously audits the contents of the various objects according to the policy set by the archive, and provides mechanisms for an independent third-party auditor to certify the integrity of any object. ACE software is developed at University of Maryland’s Institute for Advanced Computing Studies, a Chronopolis partner.

Clients may also request information regarding integrity checks of the deposited data.

Preservation Actions and Chronopolis

Please note that Chronopolis does NOT perform specific “preservation actions” upon files during or after ingest. This includes actions such as file format migration, file normalization, file type verification, creation of descriptive metadata and rights management. If a client wishes to have these services, they need to be done by the client before client Content is deposited into Chronopolis.

Chronopolis Data Access

Chronopolis is a dark preservation system. No direct access to the system will be provided to clients. Access is restricted to system administrators at each specified data center and no system administrator can access Chronopolis data at other data centers.

Chronopolis Data Retrieval

In case of critical data loss, clients can request a copy of the materials they deposited. Retrieval costs will be invoiced at $310/TB, and data is retrieved at the snapshot level. The data restoration request can be made per snapshot by an administrator using the "Request Restore" button in the DuraCloud UI. Restoration requests will begin processing within 3 working business days of initiation.

Chronopolis Data Deletion

A client may request deletion of content at the snapshot level by completing a Chronopolis Data Removal Agreement listing all snapshots to be deleted. Deletion requests will begin processing within 3 working business days of receiving a completed Removal Agreement.

Chronopolis Data Restrictions

Please note that Chronopolis prohibits deposit of works that include Personally Identifying Information (PII), Personal Health Information (PHI), or any work that includes classified information, controlled unclassified information, or any other export controlled data.

DuraCloud Health Checks and Reports

Bit integrity checks of all content are conducted twice yearly. For each content item stored in Amazon S3, the file is retrieved from storage and a checksum is calculated. This checksum is compared to both the checksum stored by S3 and the checksum maintained in the DuraCloud space manifest. Files from Amazon Glacier are not retrieved, but the checksum provided by Glacier storage is compared with the DuraCloud space manifest. The primary reason for this is cost; pulling content out of Glacier adds a significant cost overhead, which reduces its promise of being a low cost storage option. Because of this, we do not offer Glacier as a primary storage option in DuraCloud, it must be paired with S3 as primary storage.

DuraCloud provides an audit log (full list of events) and manifest for each content space listing all items and checksums. Each DuraCloud space provides a list of all content included. The DuraCloud Sync Tool provides history logs of all files transferred or updated. Additional reporting needs can be accommodated through custom development.

For content deposited in Chronopolis via DuraCloud from LYRASIS, an authorized user can retrieve a list of all files and checksums in both md5 and sha256 format.

File Replacement

If any of the bit integrity checks fail, the file is added to a failure report which is sent to the hosting provider operations staff. These staff members will re-check each failed file, and if checksums still do not match, will perform a restore action from Glacier. If the file restored from Glacier correctly matches the expected checksum, the file in S3 is replaced. If the file retrieved from Glacier also fails to match the expected checksum, we notify the customer of the discrepancy.

A subscription with a second provider, such as Glacier, allows DuraCloud to replace any items which are found to be missing or corrupt in Amazon S3. If Amazon S3 is the only provider, staff will notify the customer if a file fails integrity checks.

For content deposited in Chronopolis via DuraCloud from LYRASIS, each node in the Chronopolis network performs regularly scheduled integrity checks of all content. In the event that a file fails this test, the auditing system flags the file for review. Reviews are performed manually by Chronopolis staff in order to ascertain the cause of the audit failure. Once the cause is determined, a repair request is made to another node, which transmits a valid copy of the file for replacement at the requesting node.

Video and Audio Streaming

DuraCloud from LYRASIS hosted service spaces can be configured to enable streaming of the content stored in the space. When enabled, files can be streamed using the HLS streaming format. Streaming can be used in either open or secure modes. Secure streaming requires an authenticated request to DuraCloud to retrieve a signed URL before the stream can be delivered. More details about media streaming in DuraCloud can be found here.

Download Costs

The fees for DuraCloud integrate the cost of bandwidth and requests, allowing for downloads up to the amount of the storage subscription. What this means is that if your DuraCloud subscription is for 5 TB of content, you are able to download (retrieve) 5 TB of content each year for no additional cost. For the vast majority of DuraCloud customers, this is sufficient and there are no additional charges for download. If there is a need to download content in excess of your storage allotment, please discuss this with your hosted service staff.

Integrations

Archive-It

Archive-It partner organizations can automatically perform an offsite backup to DuraCloud, allowing independent preservation and direct access to all web archive collections captured by Archive-It.

Archivematica

DuraCloud integrates with Archivematica, and complete documentation is available. If you are interested in combining DuraCloud with a hosted Archivematica instance, you may be interested in the ArchivesDirect service.

DSpace Replication Task Suite

The DSpace Replication Task Suite is a set (suite) of tasks to assist in performing replication of DSpace content to other locations (including DuraCloud). Currently, DSpace content is packaged in containers known as archival information packages (AIPs). The DSpace Replication Task Suite was released as an optional "add-on" to DSpace 1.8 and is available in all following DSpace releases.

Fedora CloudSync

Fedora CloudSync is a web-based utility for backing up and restoring Fedora 3 content in DuraCloud. It supports on-demand and scheduled backups of any content in a Fedora 3 repository, including externally-managed datastreams. The project is functional but no longer receiving on-going support. Work on an integration between DuraCloud and Fedora 4 is currently in the planning stages.

Troubleshooting

If you encounter an error when running the SyncTool or using other content transfer tools, please first consult the list of error messages and suggested fixes below.

Reinstall the SyncTool

Often reinstalling the SyncTool will address errors or transfer issues. To reinstall the SyncTool:

1. Uninstall the SyncTool
2. Delete the work directory, called "duracloud-sync-work" in your home directory (i.e. such as C:/Users/<username>/duracloud-sync-work on Windows or /home/<username>/duracloud-sync-work on Mac/Linux)
3. Download the latest version of the SyncTool: https://wiki.lyrasis.org/display/DURACLOUD/DuraCloud+Downloads
4. Reinstall and restart your computer

Out of Memory Error or Java Heap Space Error

Please review this page for information about addressing memory errors.

SyncTool interface display problems

Restarting your machine is often all that is needed to address issues with the display of the SyncTool interface. If a restart does not address the problem please contact support.

SyncTool "file does not exist" message for a watched directory

This message will appear if one of the folders in your watched directory has been renamed. You can delete the old watched directory and add the directory with the new name.

Large files failing to upload via the SyncTool

Increasing the chunk size in the SyncTool configuration will often address this issue. Please review the SyncTool operational notes for how to optimize chunk size for larger files.

"Unable to restore file Changed List" is displayed whens starting the SyncTool

The "Changed File" is a file used by the SyncTool to keep track of the files it needs to check for possible changes. On startup the SyncTool checks this file to see if there were any files that still needed to be checked when the tool last shut down. The "Unable to restore file Changed List" message will be displayed if the SyncTool cannot read this file because it has become corrupted; this generally happens when the SyncTool did not have a chance to shut down properly, such as when a system restart occurs while the tool is running. (It is best to shut down the SyncTool prior to system reboots when possible.) Solving the problem just requires removing the Changed List file, which is in the SyncTool's work directory.

The steps to take are:

1. Stop the SyncTool (and/or check the system tray to ensure it is not running)
2. Remove the "backup" directory under the SyncTool's work directory. The work directory is named "duracloud-sync-work" and is in your home directory (i.e. such as C:/Users/<username>/duracloud-sync-work on Windows or /home/<username>/duracloud-sync-work on Mac/Linux). It is also fine to just remove the entire work directory, but this will require you to go through the setup wizard again if you're using the UI.
3. Start the SyncTool again

Technical Questions

DuraCloud is an open source project and complete user and developer documentation is publicly available.

DuraCloud Security

DuraCloud provides multiple levels of security, including an instance firewall, encrypted transmissions, application authentication, and storage provider access control.

The instance firewall provides protection to each DuraCloud instance by blocking all access except via the standard HTTP and HTTPS ports. Data transmission to and from DuraCloud is via HTTPS encrypted requests and responses that can only be read by the intended recipient. The DuraCloud application requires users accessing their DuraCloud instance via either the web or the REST API interfaces to authenticate with credentials.

Users of a DuraCloud Enterprise or Enterprise Plus instance may have various roles with associated permission levels. Users with the Administrator role have the ability to define space access controls, which defines the users and group that may read or write content in a space. Access to the underlying storage providers used by a DuraCloud instance is restricted to only DuraCloud applications. This ensures that all actions involving content must occur through DuraCloud.

DuraCloud uses the leading cloud infrastructure vendor, Amazon Web Services (AWS), for managing systems and storage. AWS has deep compliance credentials and state-of-the-art security practices. DuraCloud leverages the capabilities of AWS to deploy load balanced and auto-scaled infrastructure components which are spread across data centers to reduce localized risk exposure. Amazon S3 storage boasts 99.999999999% durability and 99.99% availability of objects over a given year.

DuraCloud is an open source technology from LYRASIS, a respected institution in the open source repository community and the home institution of the DSpace, Fedora, and VIVO communities.

Confidential Data

DuraCloud is one low-level component of an overall preservation strategy. It does not address fine-grained policy and access control considerations. DuraCloud is not audited for compliance with state or federal laws such as HIPAA or FERPA. Ensuring compliance with legal and institutional policies concerning data use and Personally-Identifying Information (PII) is the responsibility of the user/account holder. DuraCloud does provide basic authentication, space-level access controls, and the option to limit login access to a specific IP or IP range.

Encrypted Data

DuraCloud does not encrypt data. Customers may choose to encrypt files before storing them in DuraCloud, however, it is the responsibility of the customer to maintain any encryption keys.

Legal Compliance

Content access and copyright for content stored in DuraCloud is controlled and managed by the user/account holder.

Further Resources for Getting Started with Digital Preservation

We recommend The Executive Guide on Digital Preservation for learning and sharing information about digital preservation at all levels of your organization.

Below are additional resources related to digital preservation concepts, workflows, tools, and planning. With thanks to the National Digital Stewardship Alliance (NDSA) Standards & Practices interest group for creating this list.

1. National Digital Stewardship Residency (NDSR)
2. Digital Preservation Outreach and Education Network
3. DigiPres Commons
4. POWRR
5. Digital Preservation in a Box
6. NEDCC Born-Digital Preservation Reading List
7. Born-Digital & Digital Preservation Reading List
8. Digital Preservation Management Workshop
9. Digital Preservation Workflow Curriculum
10. DPC Digital Preservation Handbook
11. Digital Preservation Wikipedia project
12. Digital Preservation in iSchool Curricula
13. Society of American Archivists - Digital Archives Specialist (DAS)
14. Educopia - Sustaining Digital Curation and Preservation Training