Digital Archive – Getting Started Guide

OCLC The world's libraries. Connected.
6565 Kilgour Place, Dublin, OH 43017-3395
www.oclc.org

Contents

6565 Kilgour Place
Dublin, OH 43017-3395 USA

Introduction

This document is the user guide for collection administrators using the Digital Archive. It contains descriptions of how to use each of the three primary features of the Archive to manage your content:

1. Ingest – the process to move your content into the Archive
2. Reports – the feedback from the system you get for managing your content
3. Dissemination – the process to get copies of your content out of the Archive

Information on how to use each of these features is presented in detail in a section of this document. So that you can put these features in context, the first section is an overview describing how the Archive is organized and how it processes data.

Overview – Archive processing and organization

The Digital Archive is a managed storage environment for your content. Your content will be stored in the format and directory structure in which you send it to OCLC for processing. As each batch of content is received by OCLC, the Digital Archive performs an Ingest process. During Ingest processing, the Archive:

- checks the content against the electronic shipping manifest,
- checks each file for viruses,
- verifies each file's data format using JHOVE, and
- creates a digital fingerprint for the file so the Archive can do an independent "fixity" check to verify that no bits in the file have been altered in the future.

At the conclusion of Ingest processing, an Archive Accession Report is created for you to review.

The Digital Archive will allow any data format type to be ingested and stored. The choice of archival data format depends upon your local practice and collection policy. One purpose of the Digital Archive is for you to be able to get back the exact file you sent to OCLC for archiving.

You may include metadata files along with content files. If you do this you may want to create either a directory structure or a naming structure to associate metadata files specifically with content files. Metadata files will be treated as preservation objects like any other file.

While your data is stored in the Archive the system performs regular checks on the health of your content. The results of those checks are summarized in a File Integrity Report. You can use the results of this report to ensure that your content remains unaltered in the Archive (see Reports below).

Content in the Archive may be disseminated in two ways. You can disseminate individual files online using a web browser, or you can e-mail a request for a bulk dissemination (see Dissemination below).

Your content is organized and stored in the Archive in a structure using multiple identifiers that allow you to uniquely identify individual files and organize your collections. Each of the identifiers below is part of a hierarchy, with Institution Name at the top of the hierarchy and File Name at the bottom:

- Institution Name (OCLC symbol to be exact) – OCLC symbol is the broadest identifier for all your content. We can uniquely identify all the files in the Archive that belong to you using your OCLC symbol.
- Server URL (e.g. CONTENTdm "host name"[1]) – Identifies the source server from which content was extracted or where the access copy of the content can be found. Using Server URL allows you to group content in the Archive by the source server where associated access copies are kept.
- Collection Name (e.g. CONTENTdm "collection alias"[1]) – Identifies collections of content. This identifier is unique within a Server URL.

- Archival Volume Name (e.g. CONTENTdm "volume ID") – Identifies the batch of content you sent to OCLC. This identifier must also be unique within a Collection.
- File Name – Identifies a specific file within an archival volume. This identifier must be unique within an archival volume.

[1]This information is used in CONTENTdm servers to allow a unique link to a "preservation" file in the Archive to be tracked in the "access" file metadata associated with it.

Ingest

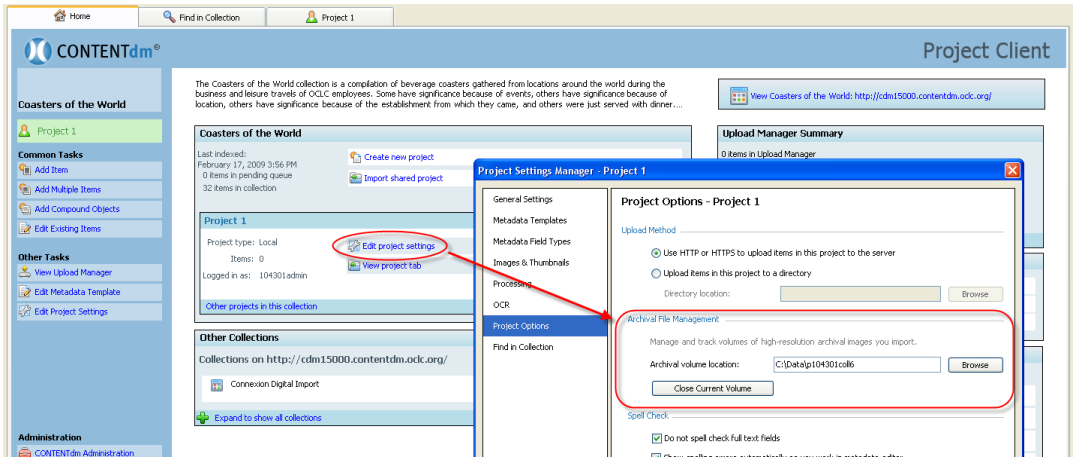
In order to move your content into the Archive, you must prepare it for submission to the Archive, and then package and ship it to the Archive.

Ingest processing in the Archive works on batches of content organized in archival volumes. An archival volume is a file system directory which contains the files you want to add to the Archive.

Prepare content for submission to the Archive – using CONTENTdm.

CONTENTdm includes an Archival File Manager that helps you prepare content for submission to the Archive. When you set up a project to load content to a collection in your CONTENTdm server, follow the directions for managing archival content at www.contentdm.com/help5/collection-admin/configuration4.asp.

These directions explain how to use the CONTENTdm Archival File Manager to prepare archival volumes on your CONTENTdm Project Client workstation. When you start a new project in Project Client, you specify the location of the archival volume in the Edit Project Settings dialog.



When you use the Archival File Manager for a project, Project Client uploads display files, thumbnail images, and metadata to the collection on your CONTENTdm server, and the original files are placed in archival volumes. These archival volumes act as a temporary staging area for your Digital Archive data.

When an archival volume is "full" the Project Client will notify you that the current volume is closed and a new volume has been created. When an archival volume is closed, an electronic shipping manifest file is completed and added to the volume. The manifest is an administrative tool used by the Archive to track, identify and manage the files in the volume. The manifest includes information about the total number of items ("ITEMTOTAL") included in that volume. The manifest includes

the following descriptive information about the files in the volume:

- Your **Institution Name** (OCLC symbol to be exact)
- Your **CONTENTdm server host name**
- The **Collection Name**
- The **Archival Volume Name**
- The **File Name**

When a volume is closed, it is ready to package and ship for ingest to the Archive.

Prepare content for submission to the Archive – without CONTENTdm.

Use these instructions if you are using a content management system other than CONTENTdm or you have content in CONTENTdm that you're not managing with Archival File Manager. To prepare content for submission to the Archive you must create an archival volume and create an electronic shipping manifest for that archival volume.

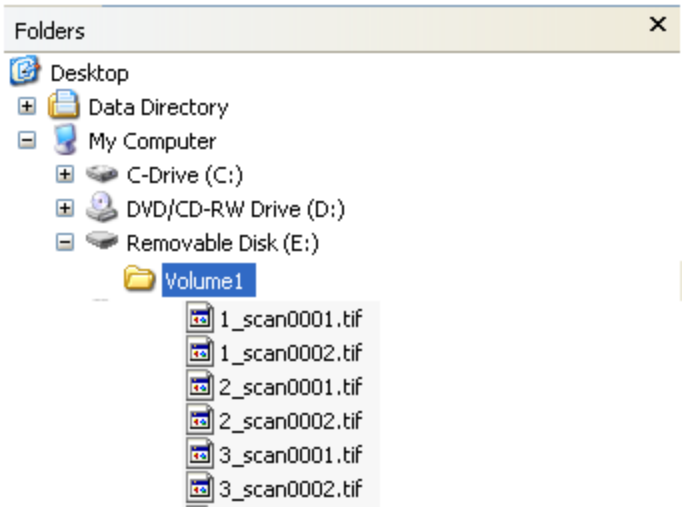
Creating an Archival Volume

Archival volumes are the package in which your data is shipped to the Archive. Using archival volumes allows you to organize multiple files and ship them to the Archive for Ingest processing.

To create an archival volume you need a "staging area", that is, space on a file system to create the archival volume directory and space within that directory to store files and create subdirectories.

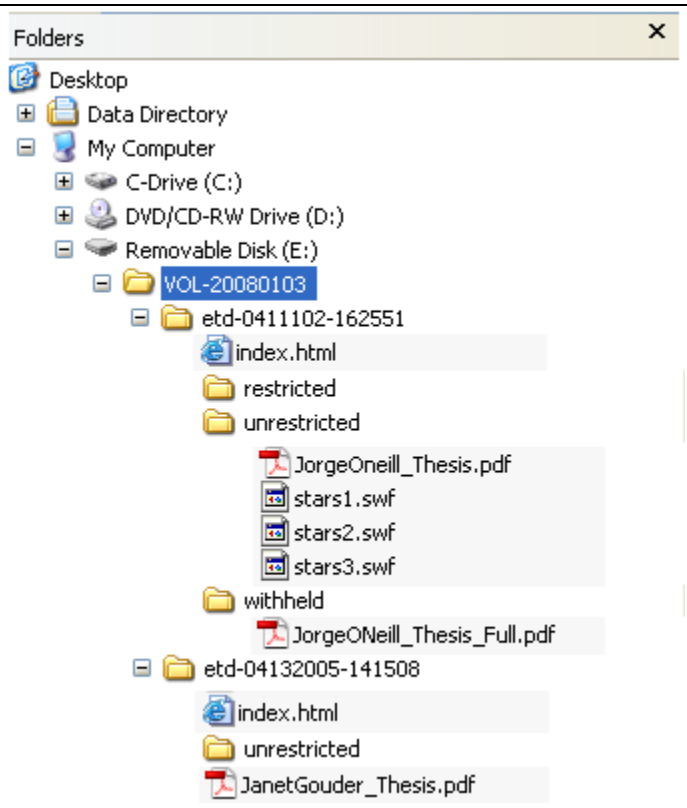
There are two types of file structures you can create in the staging area for an archival volume: single directory structure or multi-level directory structure.

A single directory structure has all the files in a single archival volume directory in the staging area with no subdirectories. Example 1 shows six TIFF files within a single archival volume named "Volume1" on a removable storage device being used as a staging area.



Example 1: single directory structure on removable disk E:

A multi-level directory structure includes subdirectories within the archival volume in the staging area. Example 2 shows various subdirectories and files within an archival volume named "VOL-20080103" on a removable storage device being used as a staging area.



Example 2: multi-level directory structure on removable disk E:

Each file in an archival volume is tracked by your OCLC symbol , your institutional URL (without the "http://"), the Collection Name, the Archival Volume Name, and the File Name. You must add all of this information to the electronic shipping manifest prior to sending it to OCLC.

Creating an Electronic Shipping Manifest

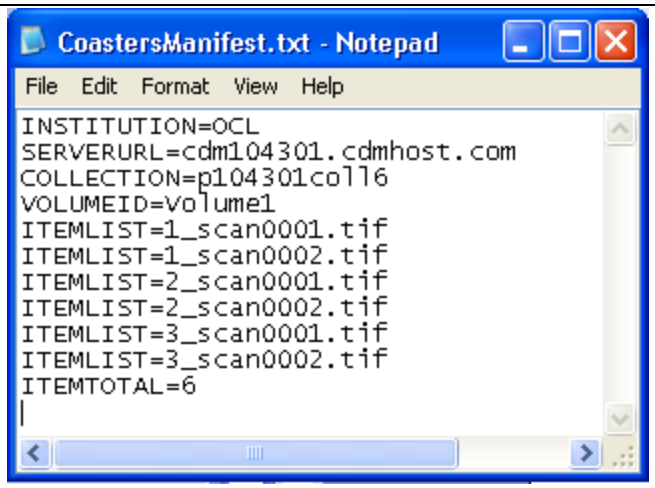
The electronic shipping manifest (manifest.txt) contains information identifying your institution and the list of files you are sending on the portable storage device. In addition, you provide information about how you want this content organized in the Archive and where to look in the directory structure of the portable storage device to find the archival volume with your content.

> **Note:** If you are using CONTENTdm Project Client to create archival volumes, an electronic shipping manifest is created and stored in your archival volume automatically (see Prepare content for submission to the Archive – using CONTENTdm. above).

There are six required data elements in the electronic shipping manifest:

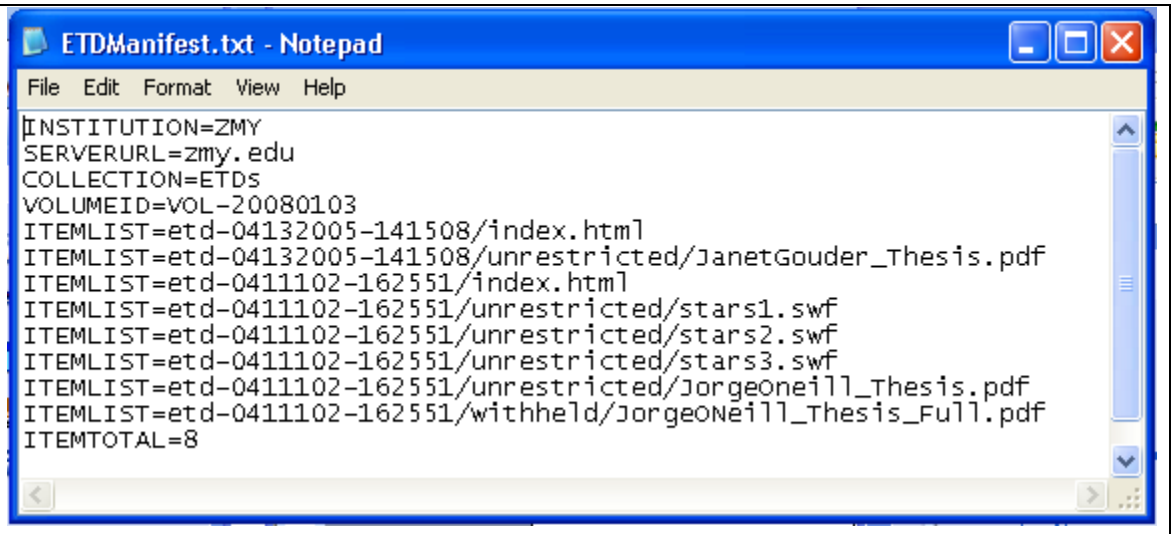
- Institution – Your institution's OCLC symbol.
- ServerURL – The full domain name of the server where corresponding access copies are kept (for example, cdm104301.cdmhost.com), or simply the web domain name of your institution (for example, **www.columbia.edu**).
- Collection – The name of the digital collection to which this content belongs.
- VolumeID – The archival volume name and also the directory name on the portable storage device where the content resides. The value of VolumeID should be unique within a collection. We recommend naming volumes with a date stamp (for example, **VOL-20080301** for a volume created on March 1, 2008).
- ItemList – The file name of a content file within the archival volume. The file name should include a mime type extension (for example, **.pdf** or **.tif**. For archival volumes with a multi-level directory structure the file name should include the relative directory path of the file within the archival volume (see example 3 below).
- ItemTotal – The count of files within the archival volume. This should be equal to the number of occurrences of the ItemList element.

For archival volumes that have a single directory structure each ItemList element contains the name and mime type extension for each file submitted.



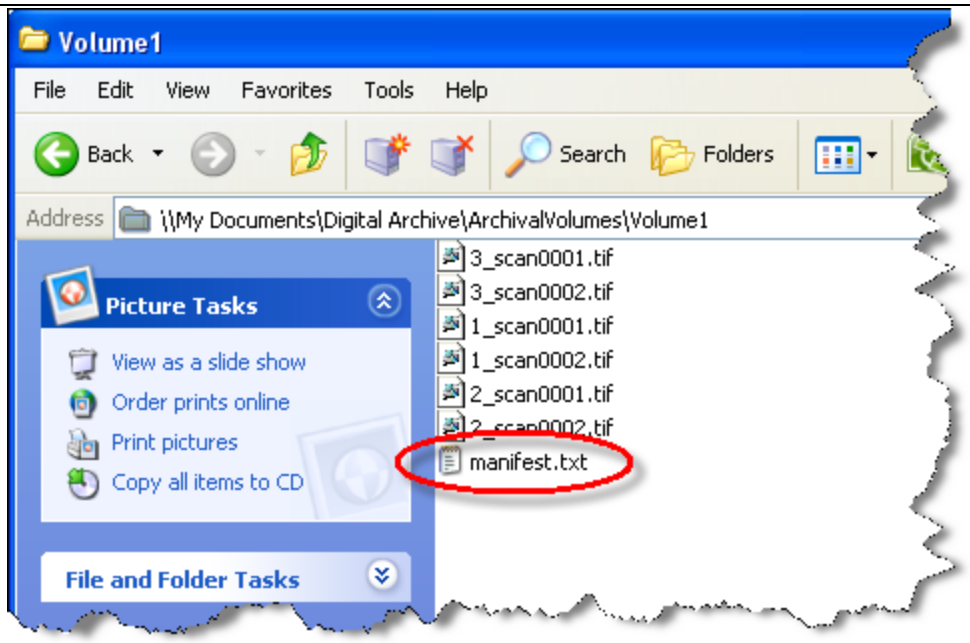
Example 3: Manifest file for a single directory structure archival volume (this is the type of archival volume created by CONTENTdm)

For archival volumes with a multi-directory structure each ItemList element contains the relative directory path of the file within the archival volume as well as the file name and mime type extension for each file submitted.



Example 4: Manifest file for a multi-directory structure archival volume

Name the electronic shipping manifest **manifest.txt** and place a copy of it inside the directory of the archival volume.



Example 5: Electronic shipping manifest inside the archival volume

Save a copy of the manifest for your records and e-mail a copy to OCLC at digitalarchive@oclc.org to notify us that you are shipping your content.

Packaging & shipping your content to the Archive

After you have prepared your archival volume and electronic shipping manifest, you must send them to the Archive for Ingest processing. The volume(s) may be shipped to the Archive on a portable storage device, or you can transfer them over the Internet to the Archive's ingest staging area.

Note: If you ship content to the Archive on a portable storage device, there are additional charges for data handling called an Ingest fee.

Transferring Archival Volumes over the Internet

The Digital Archive supports online transfer of data from your local systems to your Ingest staging area in OCLC's operations center.

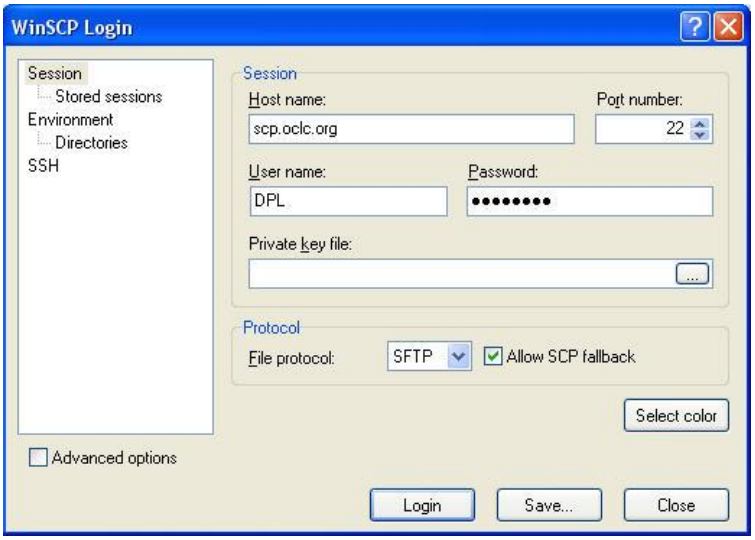
When to use this method: This method works for archival volumes less than 10 GB in total size. You need access to an FTP client such as WinSCP or FileZilla. The following directions are based on the use of WinSCP.

Before you begin … Install WinSCP: Use your browser to navigate to http://www.winscp.net/download.php. Click the link to download and save the installation package, and then run the installer to add WinSCP to your system.

Steps for sending an archival volume to the Ingest staging area using WinSCP

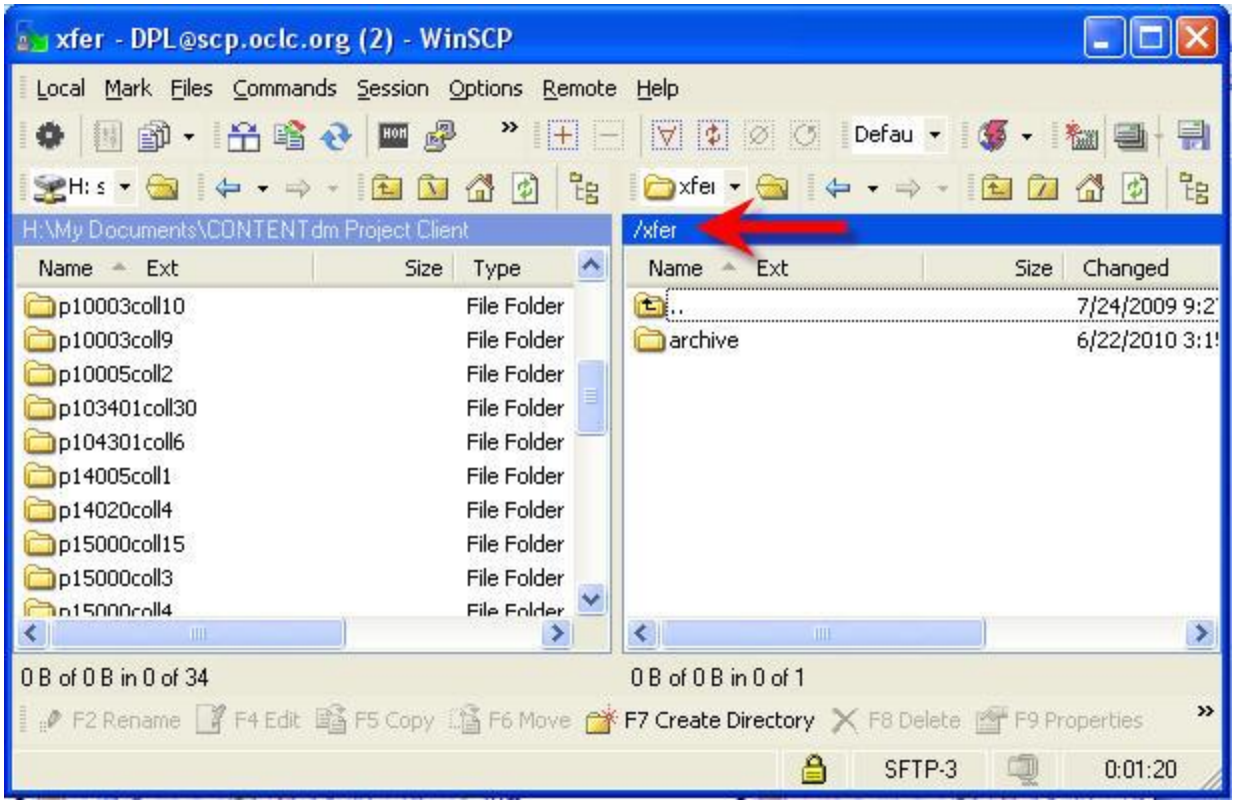
1. **Use WinSCP to login to the Digital Archive secure FTP site.**

   Open the WinSCP application and enter the host name "scp.oclc.org", plus your user name and password. Your user name is the **D.A. Uploader ID** listed in the Welcome e-mail.



   Note: If this is the first time you've used WinSCP to access scp.oclc.org, you will be asked to verify the security certificate for this site. OCLC has registered scp.oclc.org as a secure server with international Internet services, so it is OK to accept the certificate.

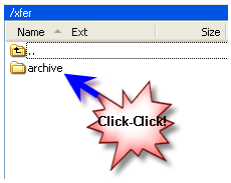
   Click **Login**, and you'll see the home directory (usually named **/xfer**) for your Digital Archive account on the right side of the display, and your workstation's local directories on the left side of the display.



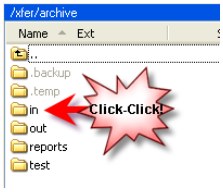
2. Change directories to the Ingest watcher directory.

Before you upload a new archival volume you need to change directories to **/xfer/archive/in/archive**. The Ingest watcher is looking for archival volumes in the **/xfer/archive/in/archive** directory.

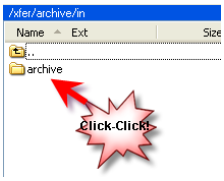
Do this by double-clicking first, on the **archive** folder …



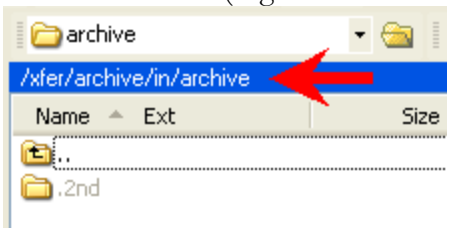
then, double-click on the **in** folder …



finally, double-click on the **archive** folder …



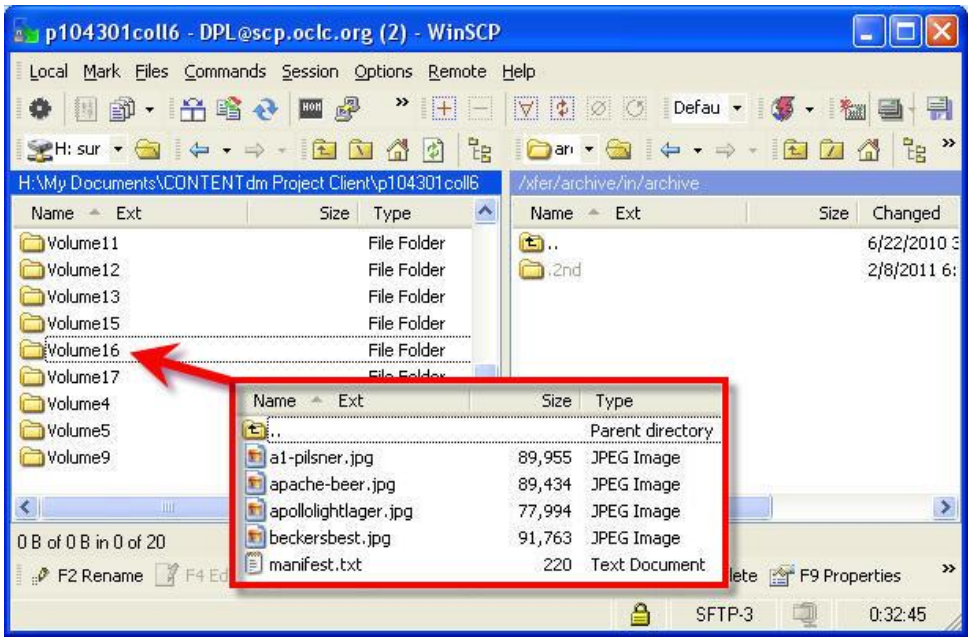
You'll end up in the **/xfer/archive/in/archive** (Ingest watcher directory)



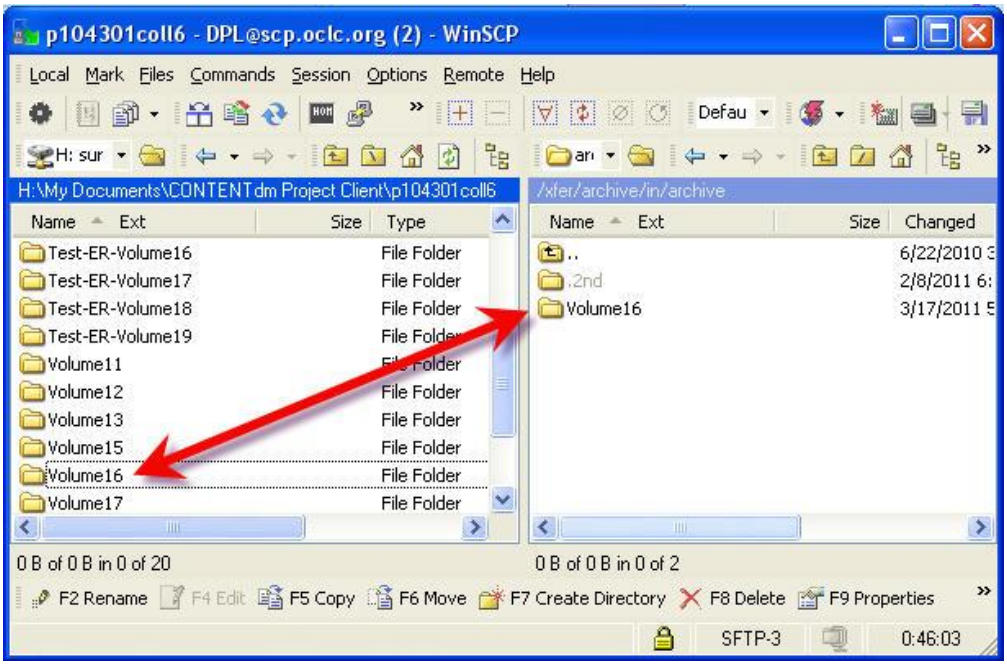
3. **Drag and drop your archival volume to the Ingest watcher directory.**

   After you've switched to **/xfer/archive/in/archive** directory on the sftp server, you are ready to copy an archival volume into this location.

   Find an archival volume on your workstation that is ready to go and includes a manifest.txt file, e.g., **Volume16** shown here.

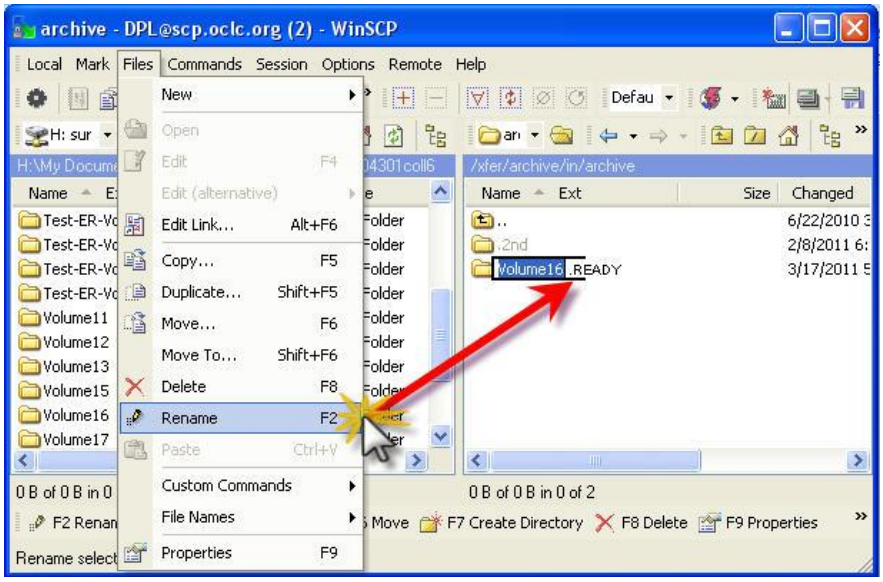


Drag and drop the archival volume from your workstation to the Ingest watcher area.



4. **Trigger the Ingest watcher.**

   Once the data transfer is complete you need to trigger the Ingest watcher. To trigger the Ingest watcher change the name of the archival volume to <name>.***READY*** using the **Files->Rename** command in WinSCP.



Remember to use ALL CAPS for the ***.READY***.

After a few minutes the file watcher will notice the .READY and start the Ingest.

If all goes well, you'll get an e-mail from the Digital Archive letting you know the success of the Ingest.

Shipping Archival Volumes on a portable storage device

You can physically ship your data to the Archive using a portable storage device.

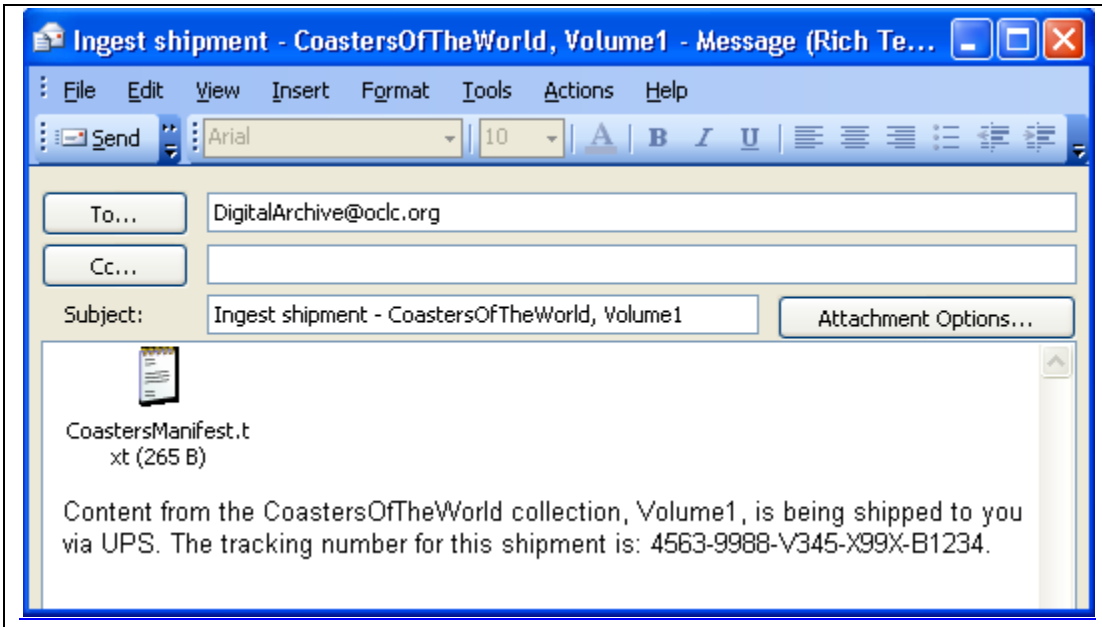
When to use this method: This method works for adding large quantities of data to the Archive, especially when you are transitioning an existing workflow to begin using the Archive.

1. After you prepare your archival volume and electronic shipping manifest as described above, copy the archival volume to a portable storage device.

2. Pack your portable storage device according to instructions from your shipping service. Address your package to the Archive:

   OCLC
   Attn: Digital Archive, MC 431
   6565 Kilgour Place
   Dublin, OH 43017

3. When you have placed your shipping order with your shipping service, send an e-mail with the electronic shipping manifest file attached to DigitalArchive@oclc.org to notify us that the package is on the way.



Example 6: E-mail notification of shipment to the Archive

Address your e-mail to DigitalArchive@oclc.org with the subject line "Ingest shipment - <collection name>, <volume ID>". Attach the electronic shipping manifest document to

the e-mail. If possible, also include your shipping service tracking number for the package in the e-mail.

Result: When your content has been received, an OCLC Archive Technician will perform Ingest processes on your archival volume and load your content into the Archive. When the Ingest completes, an Archive Accession Report will be e-mailed to you. The report also will be available in the Digital Archive Reports (see more information in the next section).

Reports

The Digital Archive produces a variety of reports to help you monitor the health of your content in the Archive. Reports for your content can be viewed at http://worldcat.org/digitalarchive/reports. You will be prompted for an authorization and password when you access this site. Use the **Reports login ID** which was sent in the welcome e-mail when you ordered the Digital Archive.

Within the Digital Archive Reports Web site, your reports are organized in folders by report type. Within a folder for a specific report type, the reports are organized chronologically by year, then by month. Within a month you will see two copies of each report generated by the system, a copy encoded in XML and a copy encoded in HTML. The HTML is easily viewable online. The XML can be used with one of the templates in the "Templates" folder to move the report information into a Microsoft Excel spreadsheet (http://worldcat.org/digitalarchive/reports/templates).

Four reports are available:

1. **Accession Report** – Produced each time new content is ingested to the Archive, this report provides a summary of Ingest processing on the content submitted to the Archive and a detailed report about the initial checking and verification performed on each file received. This report is also e-mailed to you at the completion of Ingest processing. A copy of it remains on the reports page for future reference.

2. **Activity & Storage Report** – Produced monthly, this report summarizes the entire collection and data activities, including file health, storage usage, and content accesses. This report shows the health of your Archive and changes in it.

3. **File Integrity Details Report** – Produced monthly, this report lists the details of exceptions found during routine scanning of content.

4. **Dissemination Report** – This report is produced each time a Dissemination request is processed. It includes summary information about the dissemination, including the number of files and the amount of storage. The report also includes format, virus, and fixity details about each file disseminated.

Dissemination

You can request one of two mechanisms for disseminating content from the Archive:
--Bulk dissemination requests by e-mail
Or
--Single file-by-file dissemination online using http for files smaller than 1 GB

- **About requesting Bulk Dissemination**

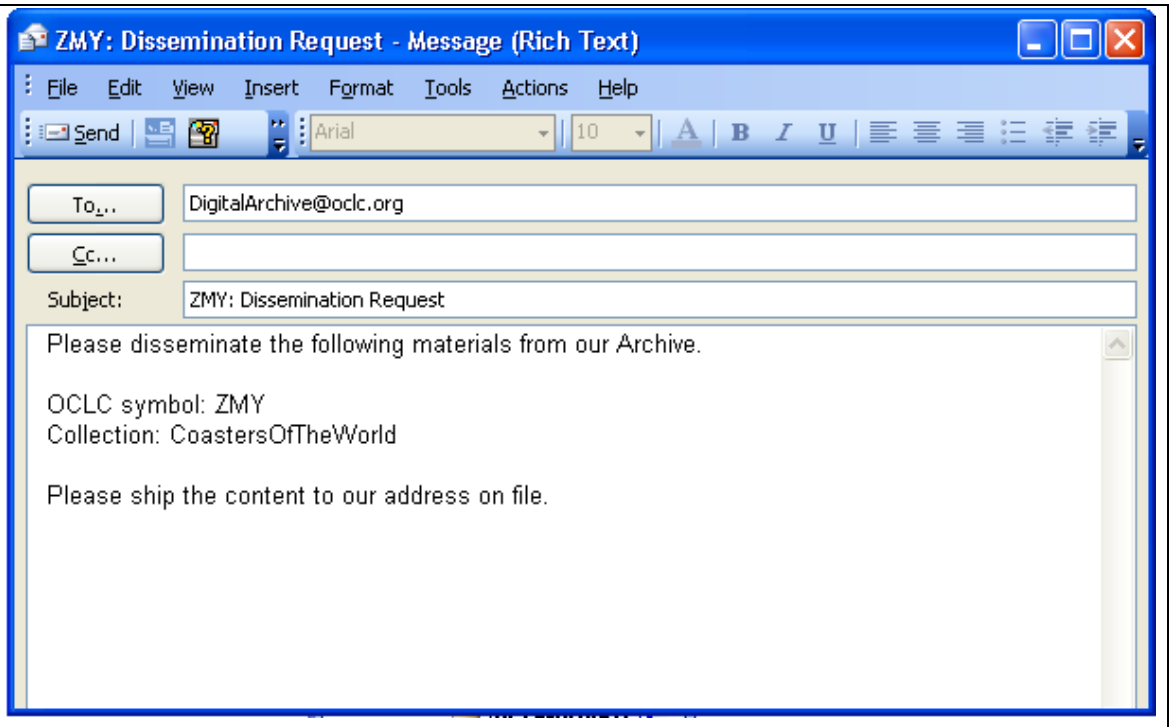
  Fees

  You are charged a Dissemination Fee and a Shipping and Media Charge for bulk dissemination requests. The Dissemination Fee allows you to request multiple disseminations in one year up to a total of one terabyte of content across all requests. Each time OCLC processes a dissemination request, you will also be charged for the portable storage device and shipping costs.

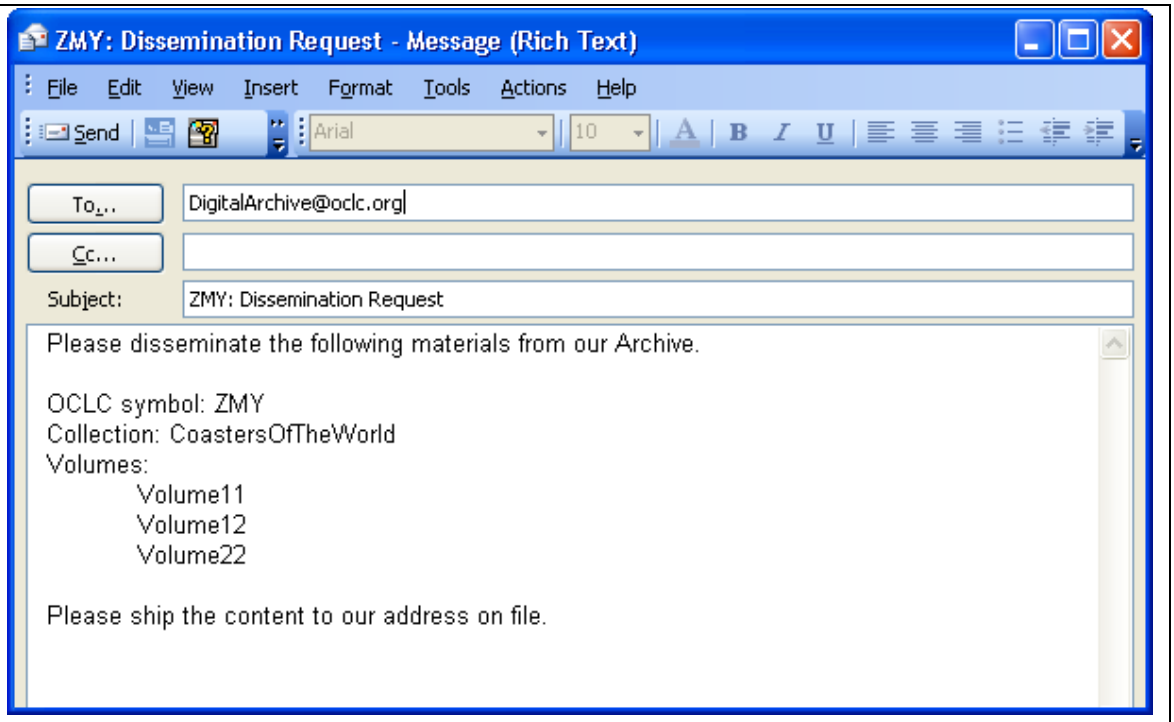
  Preparing a dissemination request
  You can request disseminations by Collection Name (see example 7 below) or by Collection Name and archival volume (see example 8 below). You can request dissemination of multiple collections or multiple archival volumes in a single e-mail request.

  To request bulk dissemination from the Archive, send an e-mail to DigitalArchive@oclc.org. Your e-mail message should include your OCLC symbol, a list of content that you want disseminated, and any special shipping instructions.

  An Archive Technician will process your request and copy the selected content to a portable storage device along with an electronic shipping manifest. At the completion of dissemination processing, the portable storage device will be shipped to you and a Dissemination Report will appear in your Digital Archive Reports Web site.



Example 7: Dissemination request – whole collection



Example 8: Dissemination request – multiple volumes in a collection

- **About requesting single file dissemination online (1 GB file size limit)**

  The Digital Archive can disseminate a single file up to 1 GB in size to a Web browser. In order to request a single file dissemination, you construct a URL based on the unique identifying elements of the file.

  Fees

  There is no fee for a single-file dissemination. These online disseminations are counted in the monthly Storage and Activity Report for your collections in the Archive.

  Configuring a single file dissemination URL

  Construct a URL for a file from these elements:

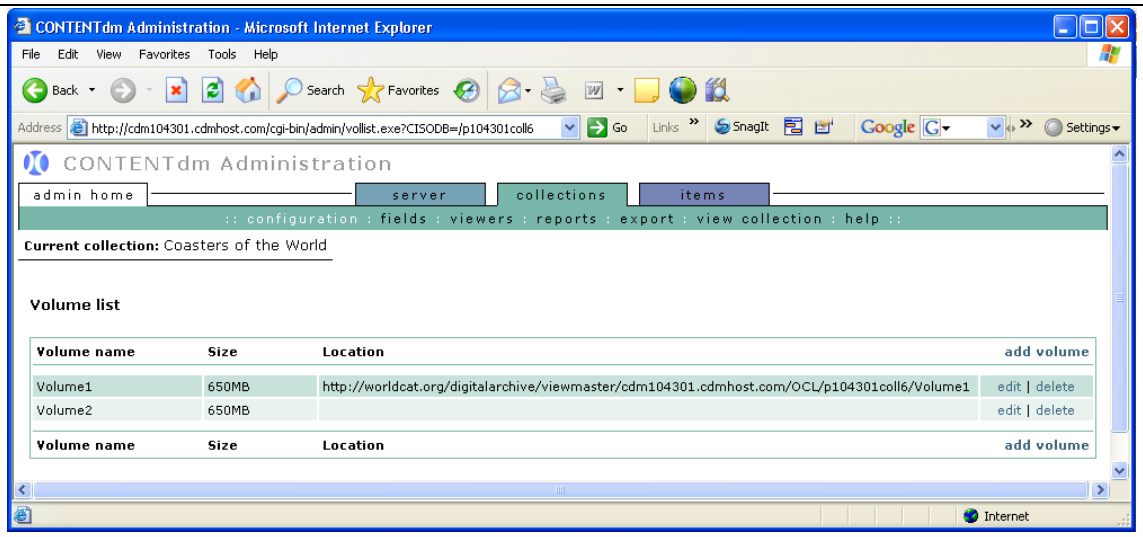
  - ROOT → http://worldcat.org/digitalarchive/viewmaster
  - Server name → /<server>[for example: /cdm104301.cdmhost.com]
  - Institution identifier → /<OCLC symbol>[for example: /ZMY]
  - Collection Name → /<collection name>[for example: /p104301coll6]
  - Volume ID → /<volume ID>[for example: /Volume1]
  - File Name → /<file name>[for example: /1_0001scan.tif]

  Altogether, using the examples shown in the list above, this URL looks like this:

  http://worldcat.org/digitalarchive/viewmaster /cdm104301.cdmhost.com/ZMY /p104301coll6/1_0001scan.tif

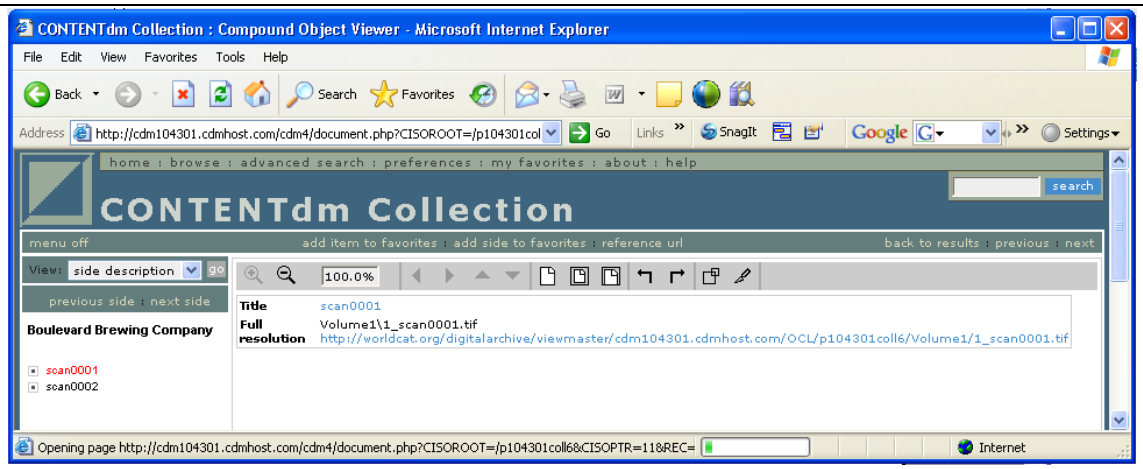
  For CONTENTdm users, the Digital Archive URL for single file dissemination can be managed in CONTENTdm Administration using Full Resolution settings and Volume Location settings. As you uploaded content to your CONTENTdm server in a collection with Full Resolution enabled, the Acquisition Station transmitted both the VolumeID and the File Name to the FullRes metadata field for **each file** uploaded. In CONTENTdm Administration, you can edit the Full Resolution Volume List "Location" field and enter the first part of the URL above (ROOT, Institution identifier, Server Name, Collection Name, VolumeID).

  Instructions for managing Volume List "Locations" are in CONTENTdm Help (see Section IV, "Moving Full Resolution Files to Permanent Locations" at http://www.contentdm.com/USC/tutorials/full-res.pdf). Entering this information once for each archival volume produces a URL for each file in the collection.



Example 9: View of CONTENTdm Administration with Volume List "Location" fields

If you edited the full resolution volume location and set the Full Res metadata field to "viewable," the metadata for the item in CONTENTdm shows a "live" hyperlink to the archived file.



Example 10: View of CONTENTdm Full Resolution Field in End-user Display

Support

The Digital Archive is monitored at all times. If you have support questions, please send an e-mail to support@oclc.org or call OCLC Customer Support at +1 800 848 5800 during the hours of 7:00 a.m. – 9:00 p.m. U.S. Eastern Time.