6/14/23
Email Archives Symposium

Preserving Email Attachments with Attachment Converter

Matt Teichman
teichman@uchicago.edu

Digital Library Development Center
Masters Program in Computer Science
University of Chicago

Nathan Mull
nmull@uchicago.edu

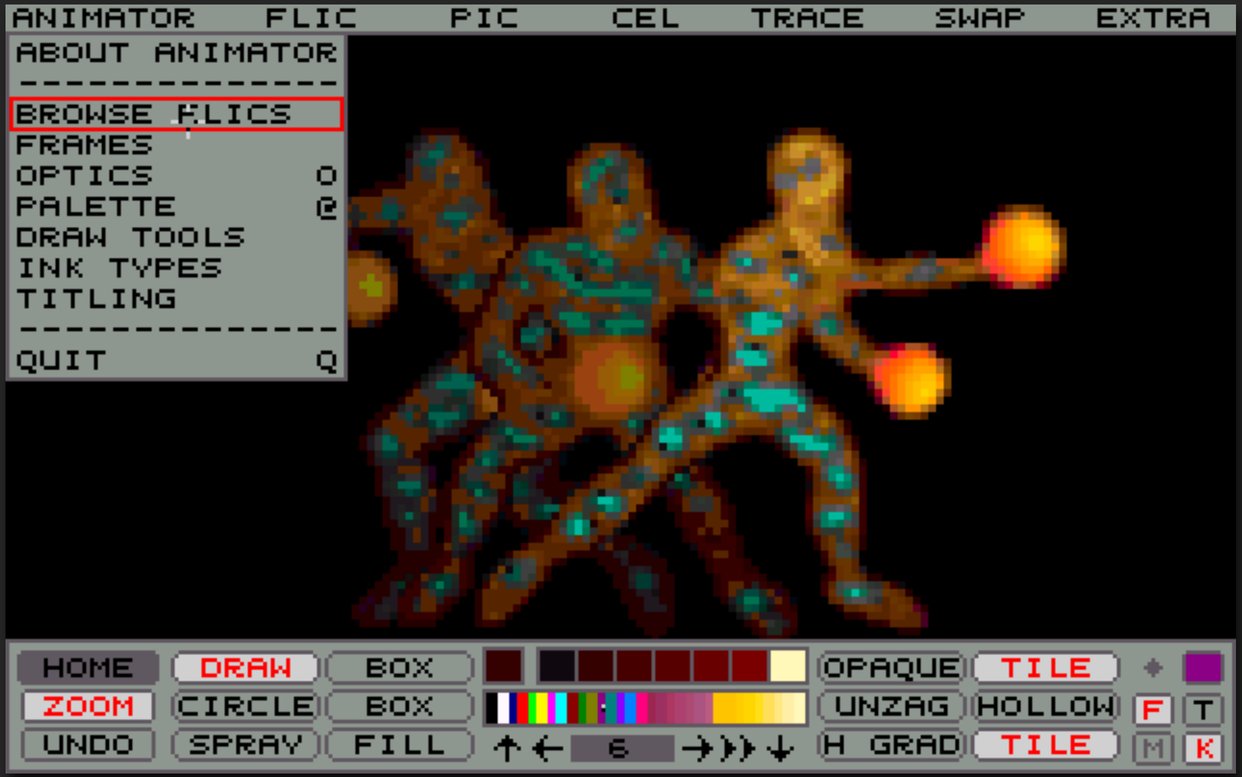
Department of Computer Science
University of Chicago

Email Attachments

Not all emails have attachments, but when an email does have an attachment it's usually some of the most important data in the entire email.

As archivists, we often have to worry about the fact that as time marches on, file formats become more and more challenging for the latest generation of computers to read.

Email Attachments



Just to take an autobiographical example, a number of the animations that Matt created when he was in high school are in formats that are challenging to open right now.

Email Attachments

What we want is for archivists of the future to be able to poke around in our email attachments and be able to open as many of them as we can.

Email Attachments

To that end, our strategy is to convert them to preservation formats today, anticipating and avoiding the future scenario in which we're stuck.

Attachment Converter

At the UChicago Digital Library Development Center, we are getting ready to release *Attachment Converter*, a tool that does exactly that.

(Thanks to the generous funding of the Email Archives: Building Capacity and Community grant, which funded the project!)

Supported Platforms

Attachment Converter is an open-source application for batch converting email attachments to preservation formats.

It runs in a Unix environment on the command line, which means you can use it on MacOS, Linux, and Windows under WSL Ubuntu.

Required Skill Level

How much computer expertise do you need to have?

Required Skill Level

You do need to be willing to open a terminal on your computer and run some commands in it to run Attachment Converter.

But you don't have to be an expert computer user beyond that to use Attachment Converter's core features.

There is one advanced feature that requires competence in UNIX system administration (more on that later), but using that feature is optional.

What Does It Do?

Attachment Converter goes through your email mailbox and does the following in batch:

- automatically converts each attachment in a format it knows about to a preservation format
- puts the converted copy of each attachment it converted back in the email it came from

Everything is Copied

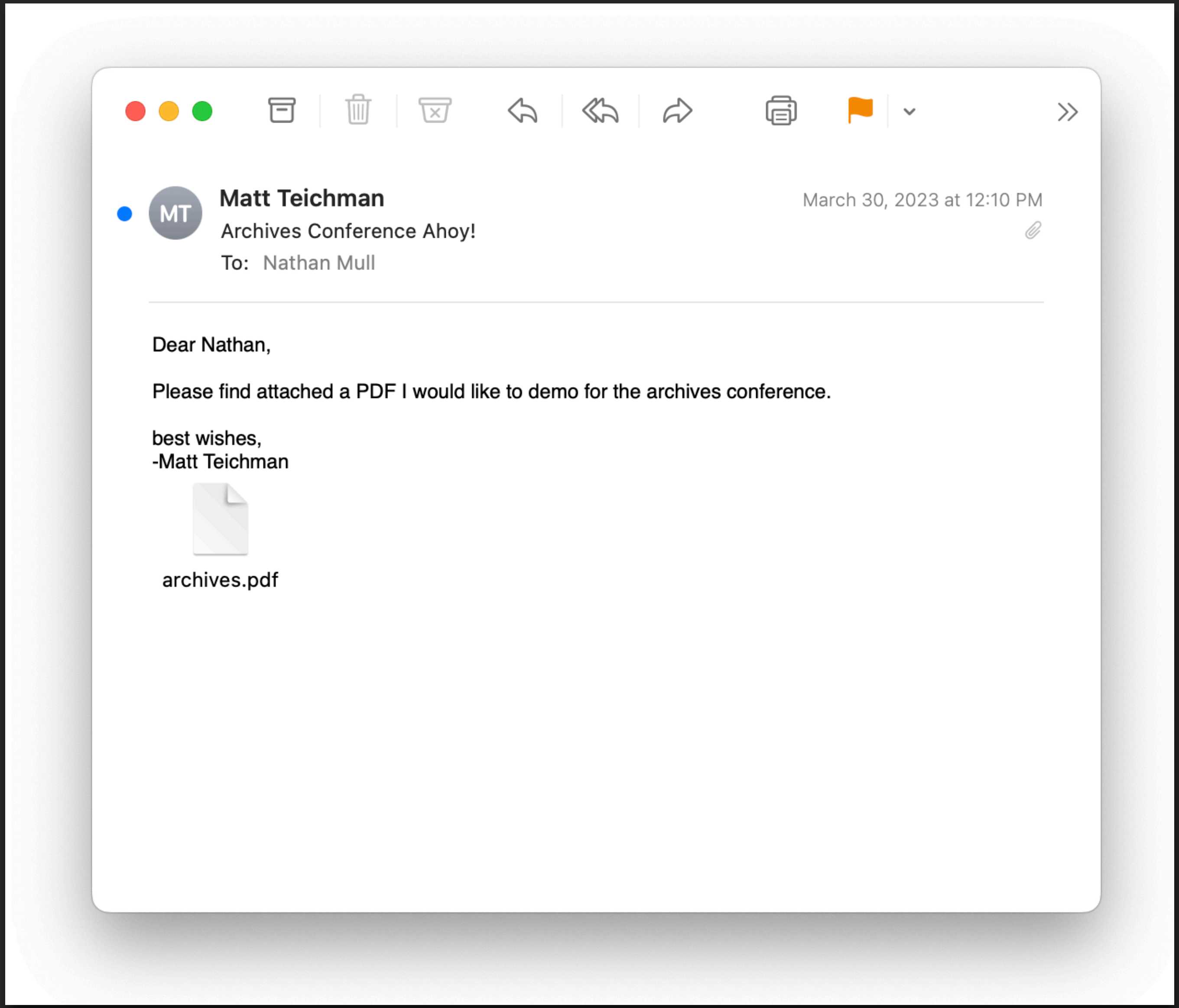
We only do *non-destructive updates*, which means that rather than converting anything in place, Attachment Converter just creates copies.

That means that:

- within each email, the converted attachment is placed next to the original, so the user has access to both
- when you convert a mailbox, the original is left alone while Attachment Converter creates a copy of the entire mailbox

Everything is Copied

So if this was one of the original email in your mailbox:



Matt Teichman

Archives Conference Ahoy!

To: Nathan Mull

March 30, 2023 at 12:10 PM

Dear Nathan,

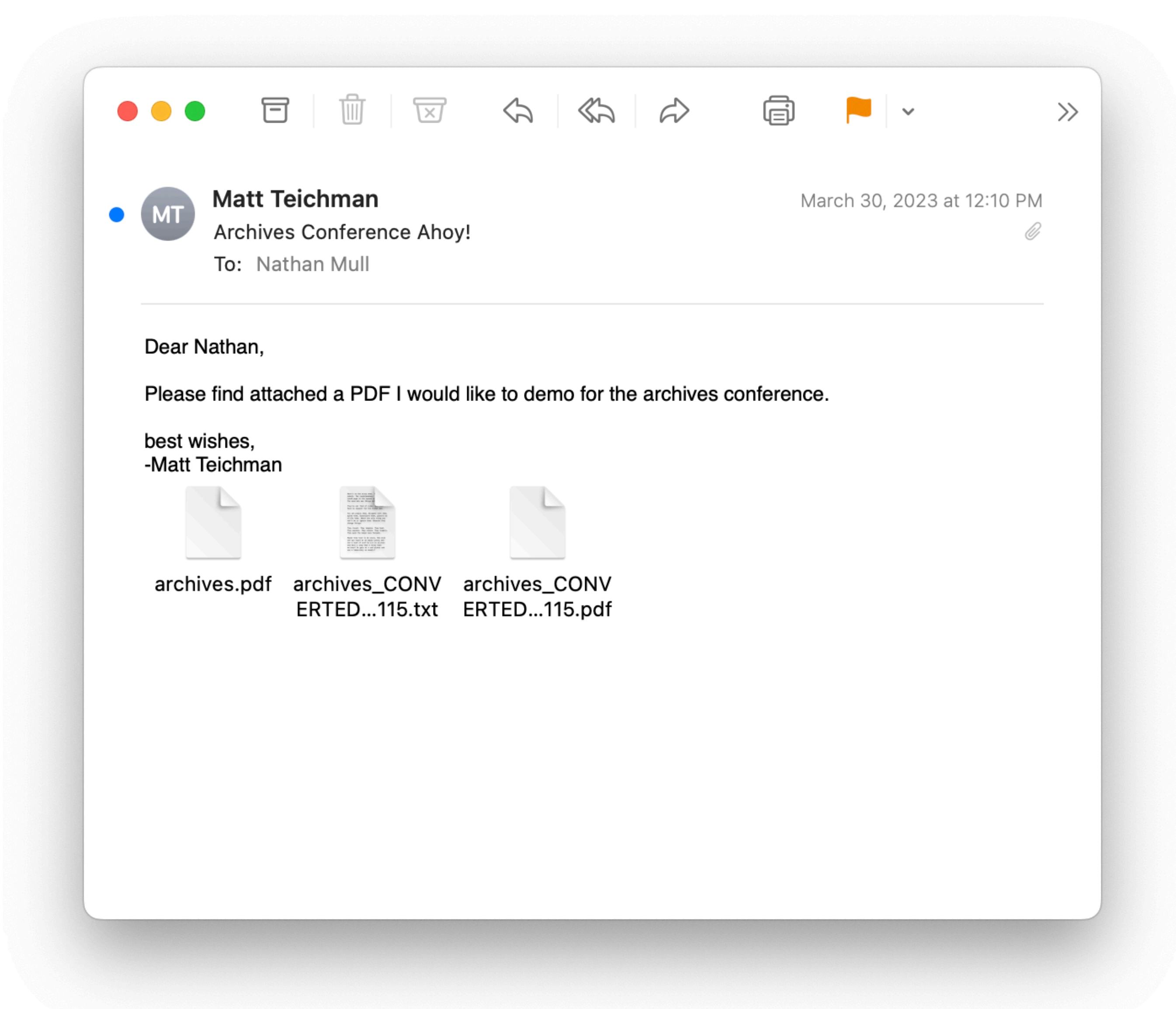
Please find attached a PDF I would like to demo for the archives conference.

best wishes,
-Matt Teichman

archives.pdf

Everything is Copied

Attachment Converter would make this email as a copy:



Matt Teichman

Archives Conference Ahoy!

To: Nathan Mull

March 30, 2023 at 12:10 PM

Dear Nathan,

Please find attached a PDF I would like to demo for the archives conference.

best wishes,
-Matt Teichman

Original Order

What we like about this approach is that it obeys the archival principles of *provenance* and *original order*:

- the attachment is preserved in a location that shows who was sending it to whom and why
- the order in which copied attachments appear corresponds to the order in which they were sent

We use MBOX

The mailbox format Attachment Converter uses is MBOX, which is the format provided by GMail and Apple Mail for export.

MBOX has three characteristics that make it particularly great for use in archives:

- it has withstood the test of time
- it is widely supported
- it is human-readable

Conversion Outsourced

Attachment Converter does not do any of the file format conversions itself.

Instead, it can invoke whatever external utilities for converting one file to another are already installed on your computer.

(If you don't have any such utilities, our installation package will install some for you.)

Batch Conversion

What Attachment Converter *does* do is:

- find the attachments in a mailbox that are in a given format
- convert them to whatever format the user would like
- pack the converted copies back into the original email

Batch Conversion

By default, Attachment Converter will perform the following conversions:

- PDF → PDF-A
- PDF → plaintext
- DOC/X → PDF-A
- DOC/X → plaintext
- XLS/X → PDF-A
- XLS/X → TSV
- JPG → TIFF
- PNG → TIFF
- GIF → TIFF

Advanced Features

More ambitious users can also take advantage of Attachment Converter's advanced features, which allow you to use any conversion utility you want to convert attachments.

This involves editing Attachment Converter's configuration file to tell it what file format to look for, what file format to convert to, and where on your computer the program that will do the conversion is located.

Config File

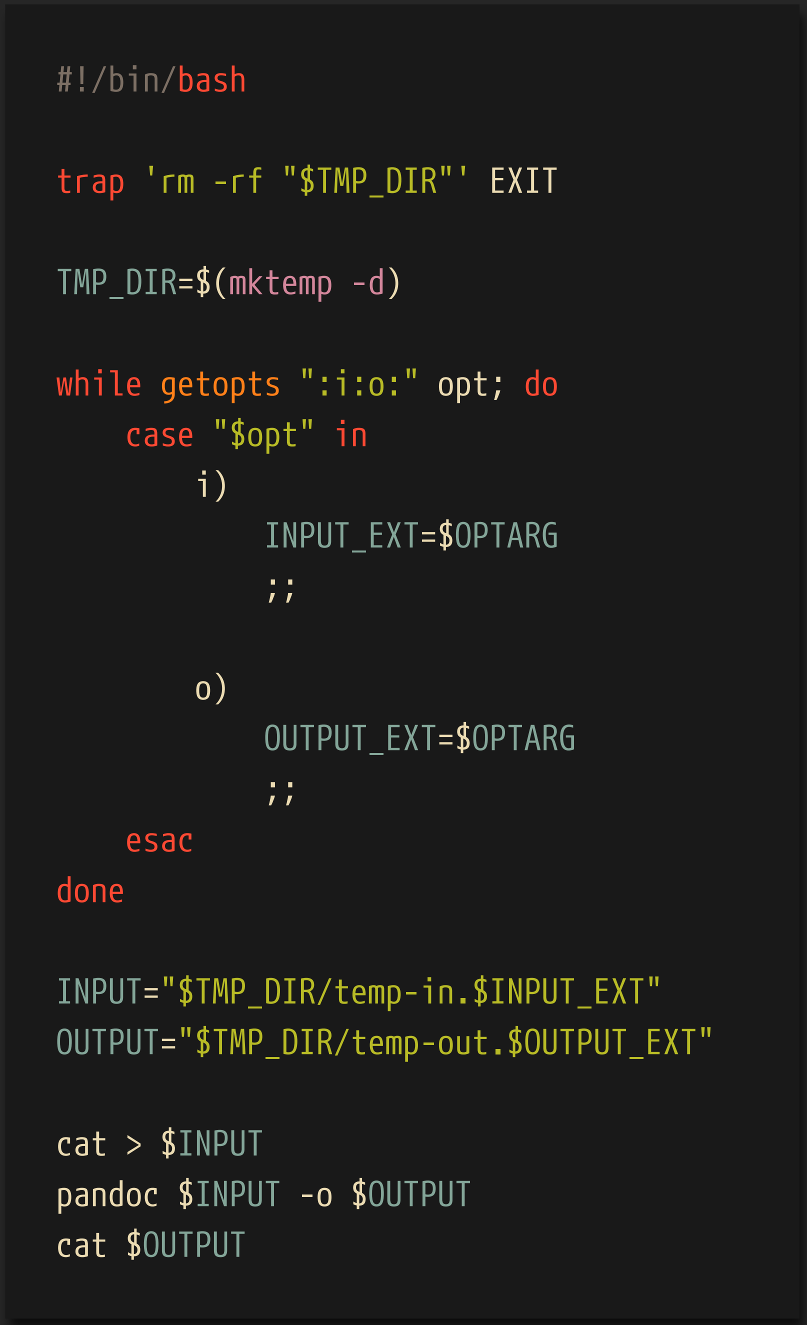
The configuration file looks (more or less) like this.

```
%source_type application/pdf
%target_type text/plain
%shell_command .../pdftotext-wrapper.sh
%id pdftotext-pdf-to-text

%source_type application/msword
%target_type application/pdf
%shell_command .../soffice-wrapper.sh -i doc -o pdf
%id soffice-doc-to-pdfa

%source_type application/msword
%target_type text/plain
%shell_command .../soffice-wrapper.sh -i doc -o txt
%id soffice-doc-to-txt
```

Advanced Features



```
#!/bin/bash

trap 'rm -rf "$TMP_DIR"' EXIT

TMP_DIR=$(mktemp -d)

while getopts ":i:o:" opt; do
    case "$opt" in
        i)
            INPUT_EXT=$OPTARG
            ;;

        o)
            OUTPUT_EXT=$OPTARG
            ;;
    esac
done

INPUT="$TMP_DIR/temp-in.$INPUT_EXT"
OUTPUT="$TMP_DIR/temp-out.$OUTPUT_EXT"

cat > $INPUT
pandoc $INPUT -o $OUTPUT
cat $OUTPUT
```

Customizing Attachment Converter also involves writing a shell script to make sure the external conversion utility you choose to use follows the specification of a UNIX filter.

Advanced Features

If you don't know what those words mean, don't worry!

- you can enlist the aid of anyone who is familiar with UNIX system administration to help you with this advanced configuration

- and you can always just use Attachment Converter with its default configuration–this is just for if you have a custom file format conversion you want to do

Suggestions

Another perk to the project being in its early stages is that you can always email us if you have suggestions about new features for us to add or new formats to support:

Matt Teichman	*Nathan Mull*
teichman@uchicago.edu	*nmull@uchicago.edu*

Later This Year

Attachment Converter is due for release in late summer / early fall of 2023.

Please check our website for updates:

`https://dldc.lib.uchicago.edu/open/attachment-converter`