Apache Airflow Blinken OSA case study

JÓZSEF GÁBOR BÓNÉ - HEAD OF IT
BONEJ@CEU.EDU

GITHUB.COM/BLINKENOSA/WORKFLOWS

Docker containers

postgres:9.6

database server for permanent storage

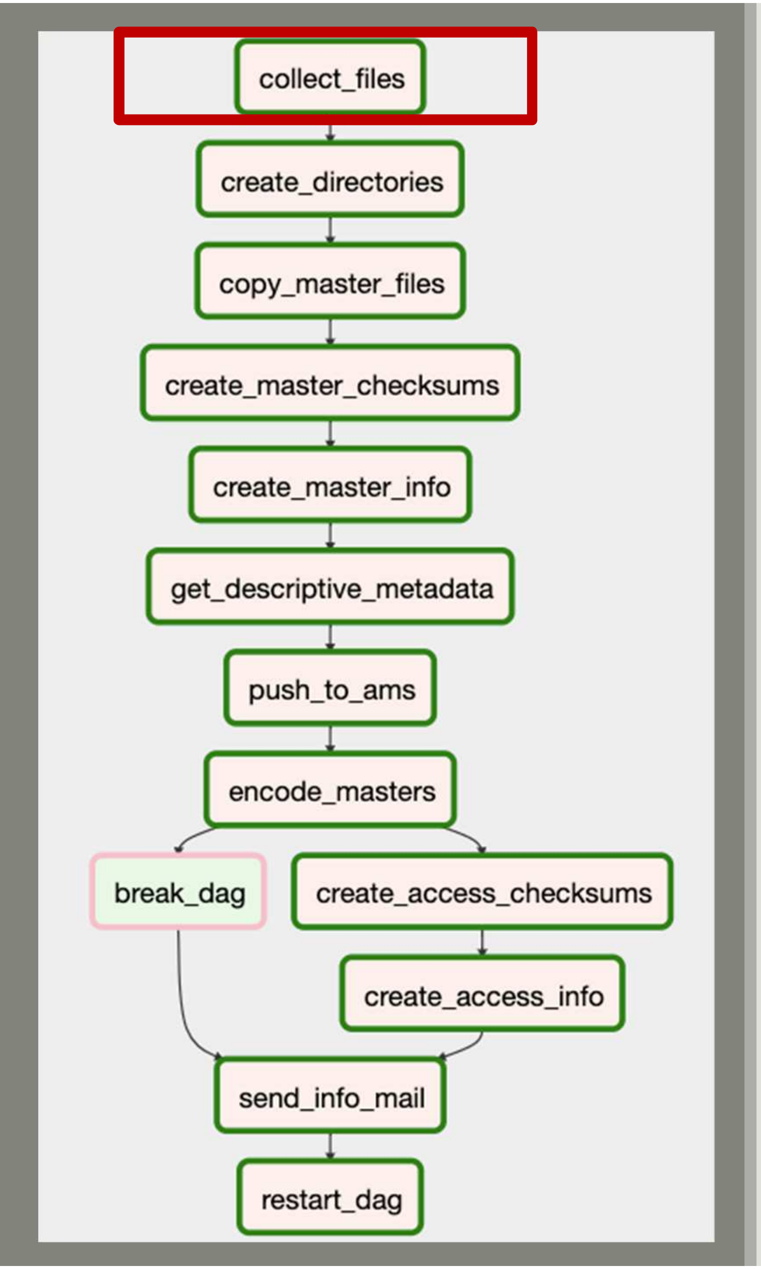
puckel/docker-airflow

apache airflow in a docker container

nightseas/ffmpeg

ffmpeg with compiled CUDA drivers*

* server is equipped with NVIDIA GTX 1080



Preservation workflow for video files - steps

collect_files

Input

Directory with filenames in the format of OSA barcrode.
Example: `HU_OSA_00000011.avi`

OR

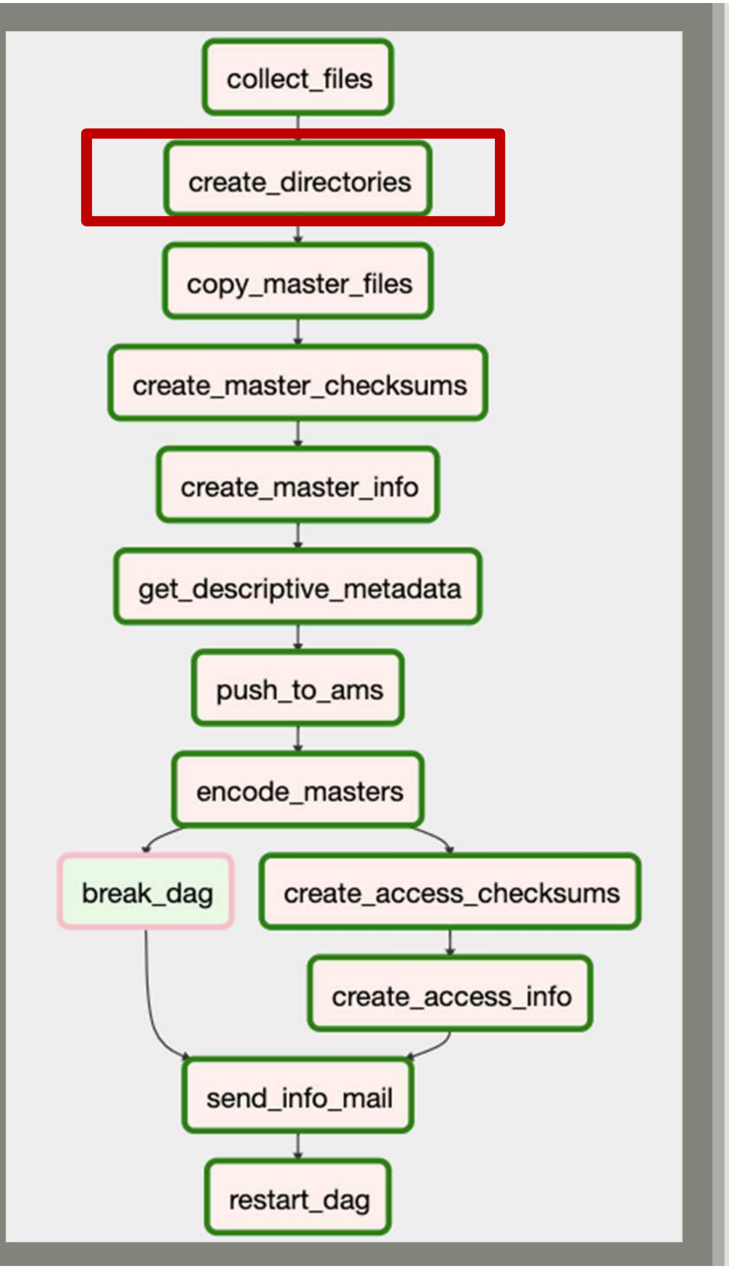
Directories with names in the format of OSA barcode.
Example: `HU_OSA_00000011/movie.avi`

Task

Picks up the first file from the input directory, moves it to a working directory and places the name and location into 'videofiles.json' file.

Example:

```
{'HU OSA 00000011': '/opt/videos/av_hdd/HU_OSA_00000011.avi'}
```



Preservation workflow for video files - steps

create_directories

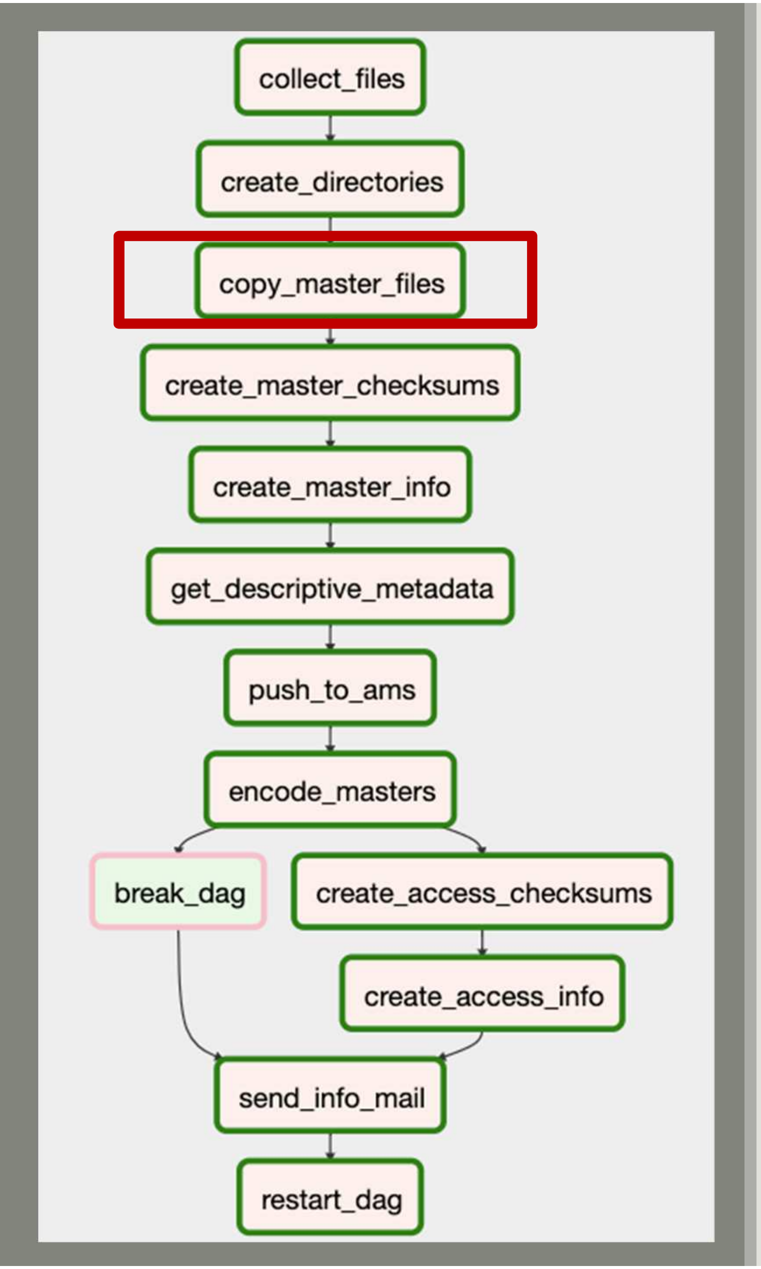
<u>Task:</u>

Create directory structure for the AIP.

Example:

```
HU_OSA_00000011
    Content
        Access
        Preservation
    Metadata
        Access
        Preservation
```

Preservation workflow for video files - steps



copy_master_files

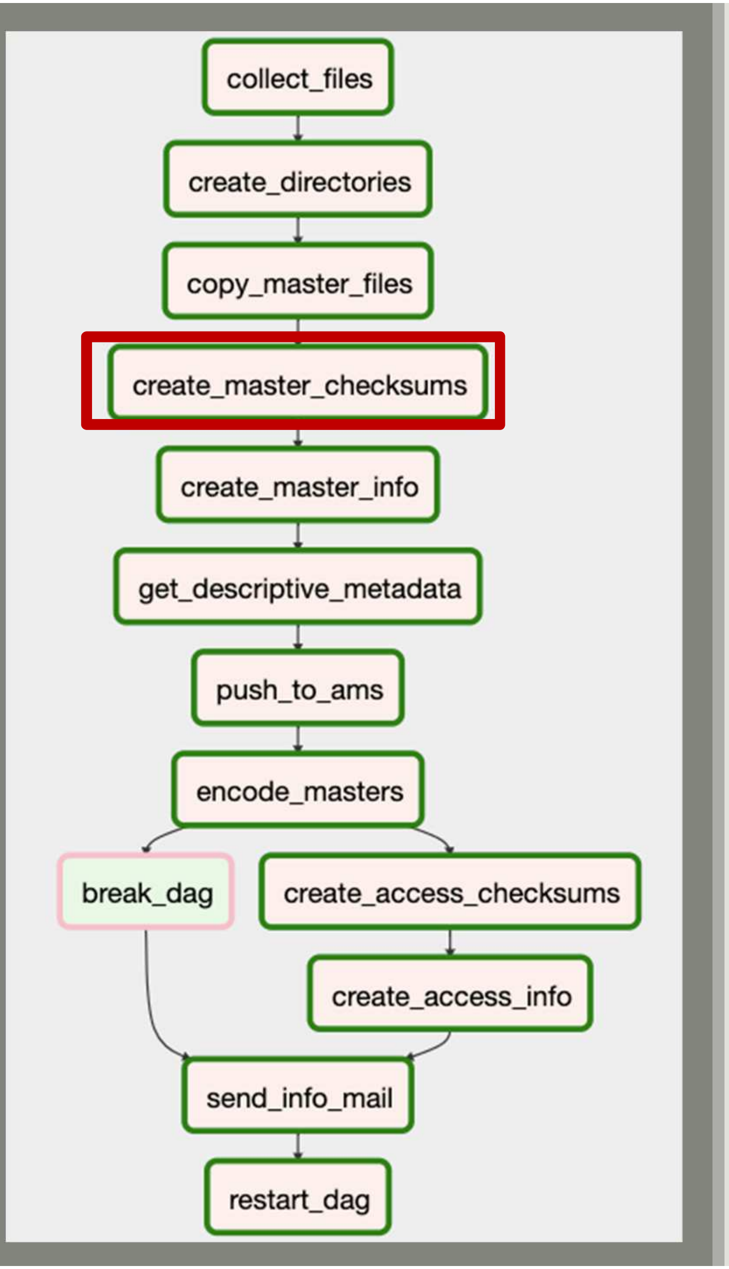
<u>Task:</u>

Move master file to the appropriate directory.

Example:

```
HU_OSA_00000011
    Content
        Access
        Preservation
            HU_OSA_00000011.avi
    Metadata
        Access
        Preservation
```

Preservation workflow for video files - steps



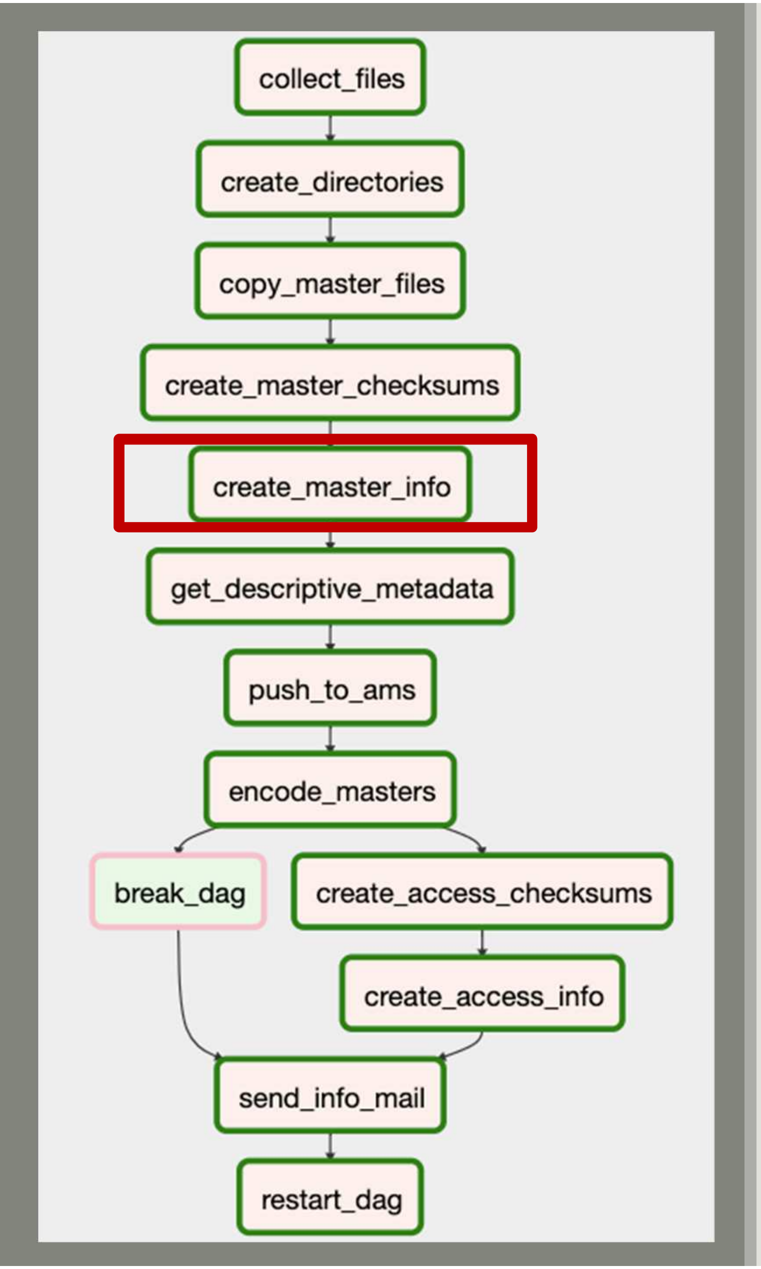
create_master_checksums

Task:

Create md5 and sha512 checksums for the master file.

Example:

```
HU_OSA_00000011
    Content
        Access
        Preservation
            HU_OSA_00000011.avi
    Metadata
        Access
        Preservation
            HU_OSA_00000011.md5
            HU_OSA_00000011.sha512
```



Preservation workflow for video files - steps

create_master_info

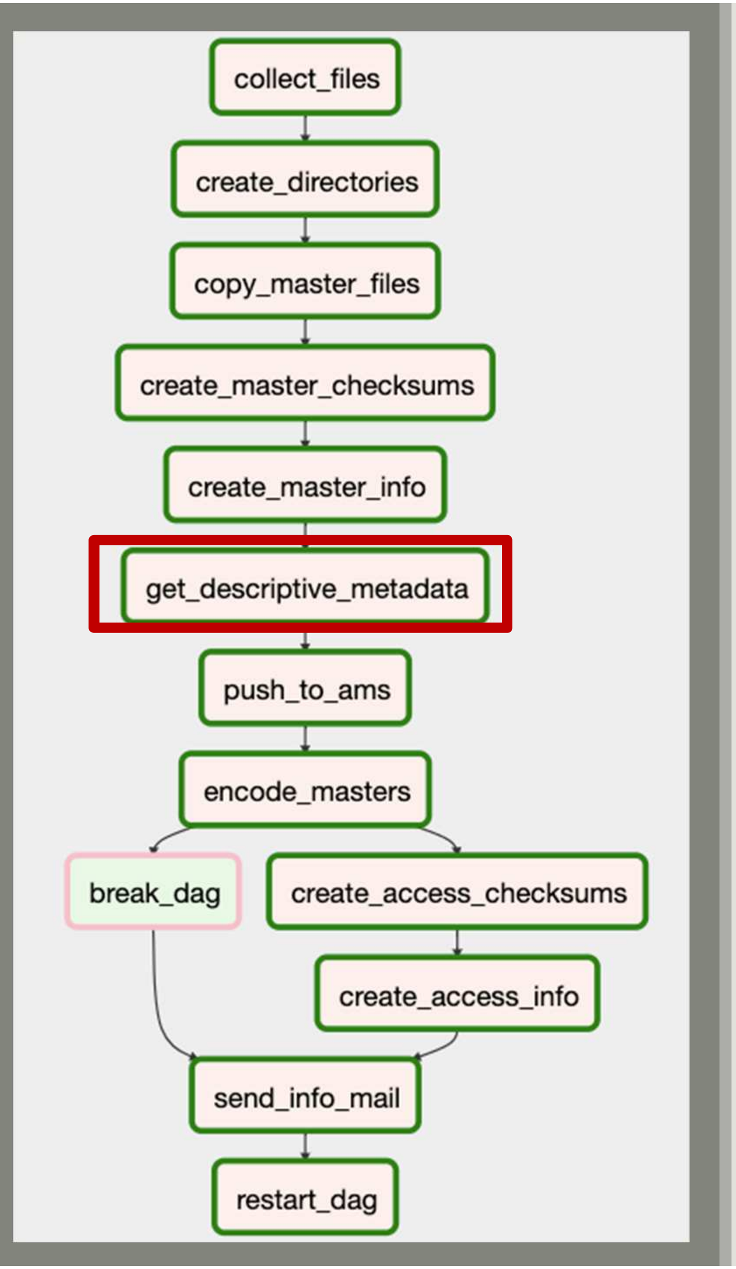
<u>Task:</u>

Save the output of 'ffprobe' command for master.

Example:

```
HU_OSA_00000011
    Content
        Access
        Preservation
            HU_OSA_00000011.avi
    Metadata
        Access
        Preservation
            HU_OSA_00000011.md5
            HU_OSA_00000011.sha512
            HU_OSA_00000011_md_tech.json
```

Preservation workflow for video files - steps



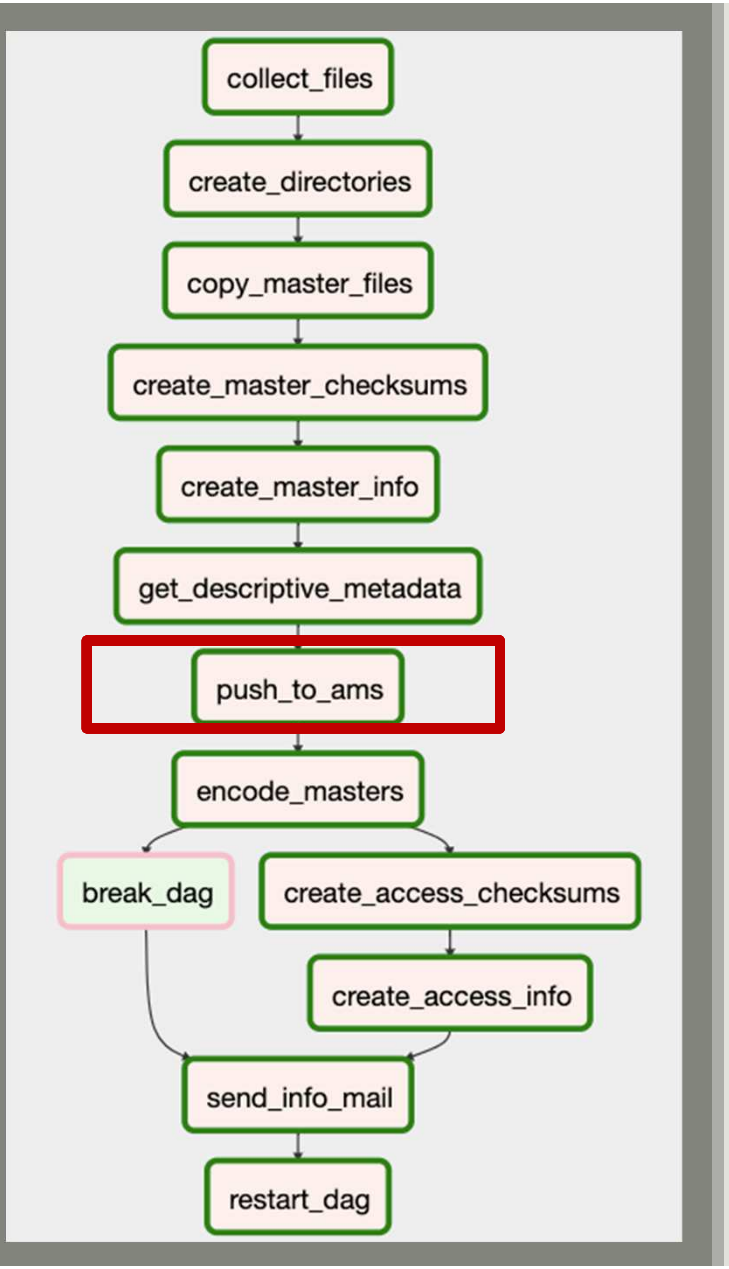
get_descriptive_metadata

Task:

If exists save descriptive metadata by querying the API (HTTP GET) of the Archival Management System.

Example:

```
HU_OSA_00000011
    Content
        Access
        Preservation
            HU_OSA_00000011.avi
    Metadata
        Access
        Preservation
            HU_OSA_00000011.md5
            HU_OSA_00000011.sha512
            HU_OSA_00000011_md_descriptive.json
            HU_OSA_00000011_md_tech.json
```



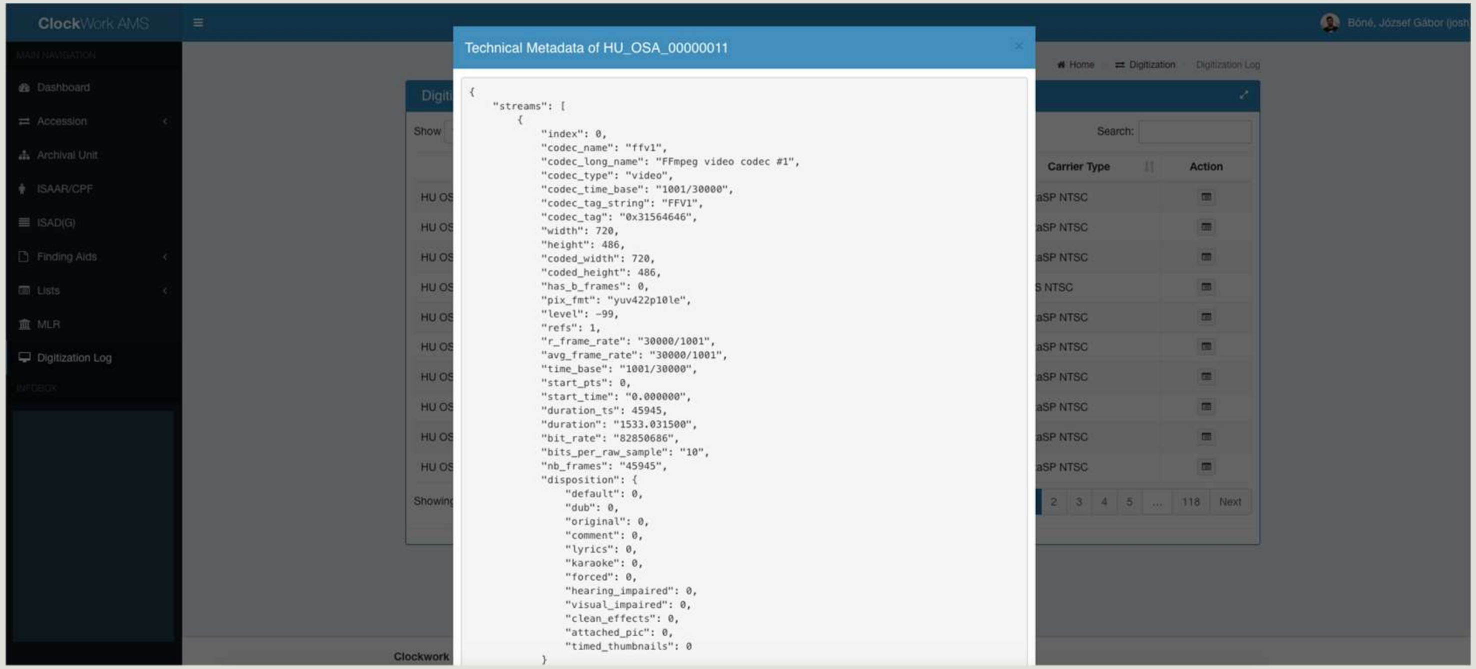
Preservation workflow for video files - steps

push_to_ams

Task:

Save technical metadata in the Archival Management System by submitting it to the API (HTTP POST).

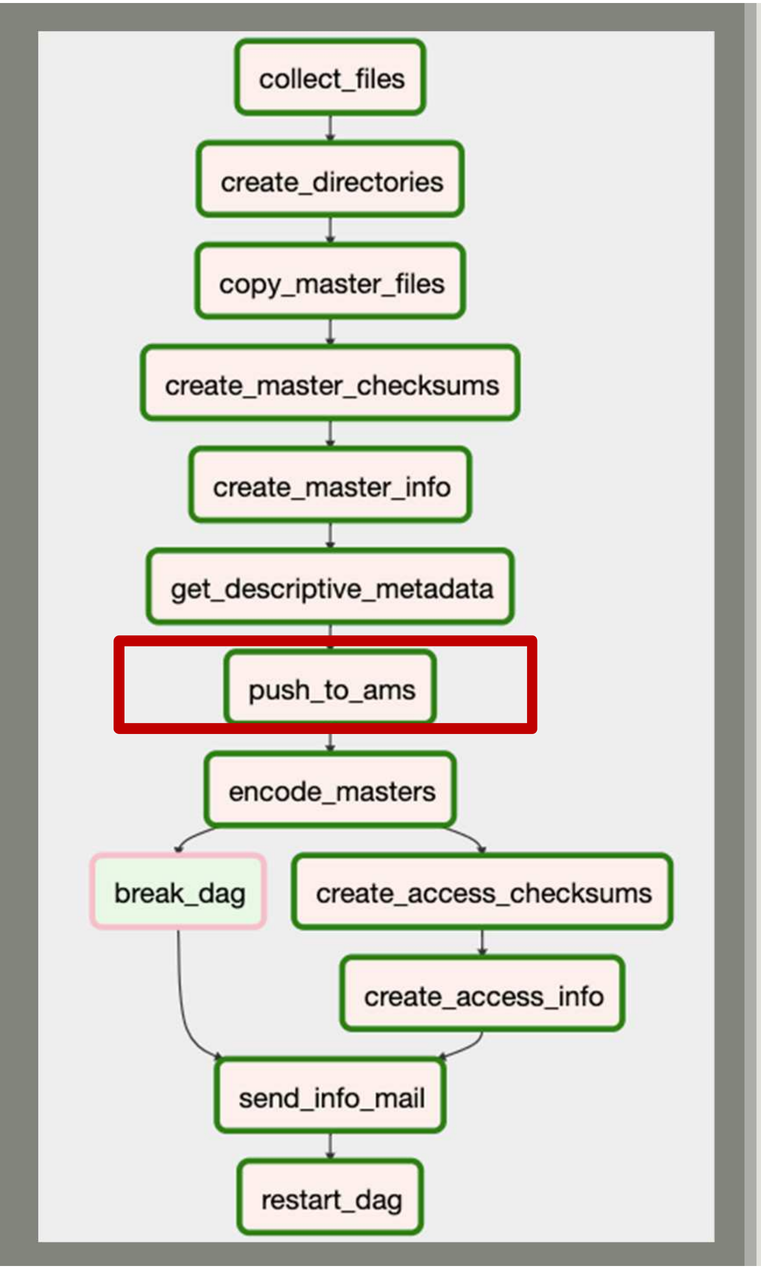
Example:



Technical Metadata of HU_OSA_00000011

```
{
    "streams": [
        {
            "index": 0,
            "codec_name": "ffv1",
            "codec_long_name": "FFmpeg video codec #1",
            "codec_type": "video",
            "codec_time_base": "1001/30000",
            "codec_tag_string": "FFV1",
            "codec_tag": "0x31564646",
            "width": 720,
            "height": 486,
            "coded_width": 720,
            "coded_height": 486,
            "has_b_frames": 0,
            "pix_fmt": "yuv422p10le",
            "level": -99,
            "refs": 1,
            "r_frame_rate": "30000/1001",
            "avg_frame_rate": "30000/1001",
            "time_base": "1001/30000",
            "start_pts": 0,
            "start_time": "0.000000",
            "duration_ts": 45945,
            "duration": "1533.031500",
            "bit_rate": "82850686",
            "bits_per_raw_sample": "10",
            "nb_frames": "45945",
            "disposition": {
                "default": 0,
                "dub": 0,
                "original": 0,
                "comment": 0,
                "lyrics": 0,
                "karaoke": 0,
                "forced": 0,
                "hearing_impaired": 0,
                "visual_impaired": 0,
                "clean_effects": 0,
                "attached_pic": 0,
                "timed_thumbnails": 0
            }
```

Preservation workflow for video files - steps

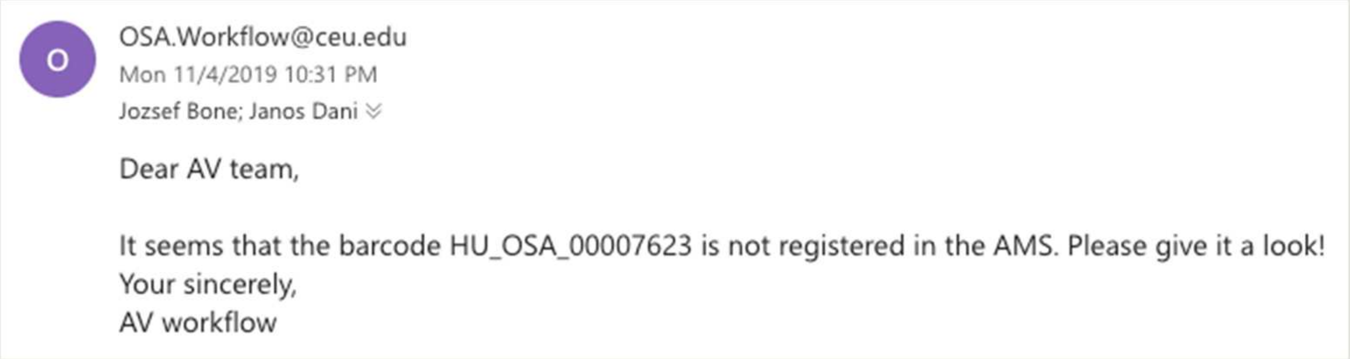


push_to_ams

<u>Error:</u>

If the barcode can't be found in the AMS, OSA AV digitization staff gets an email with a warning message.

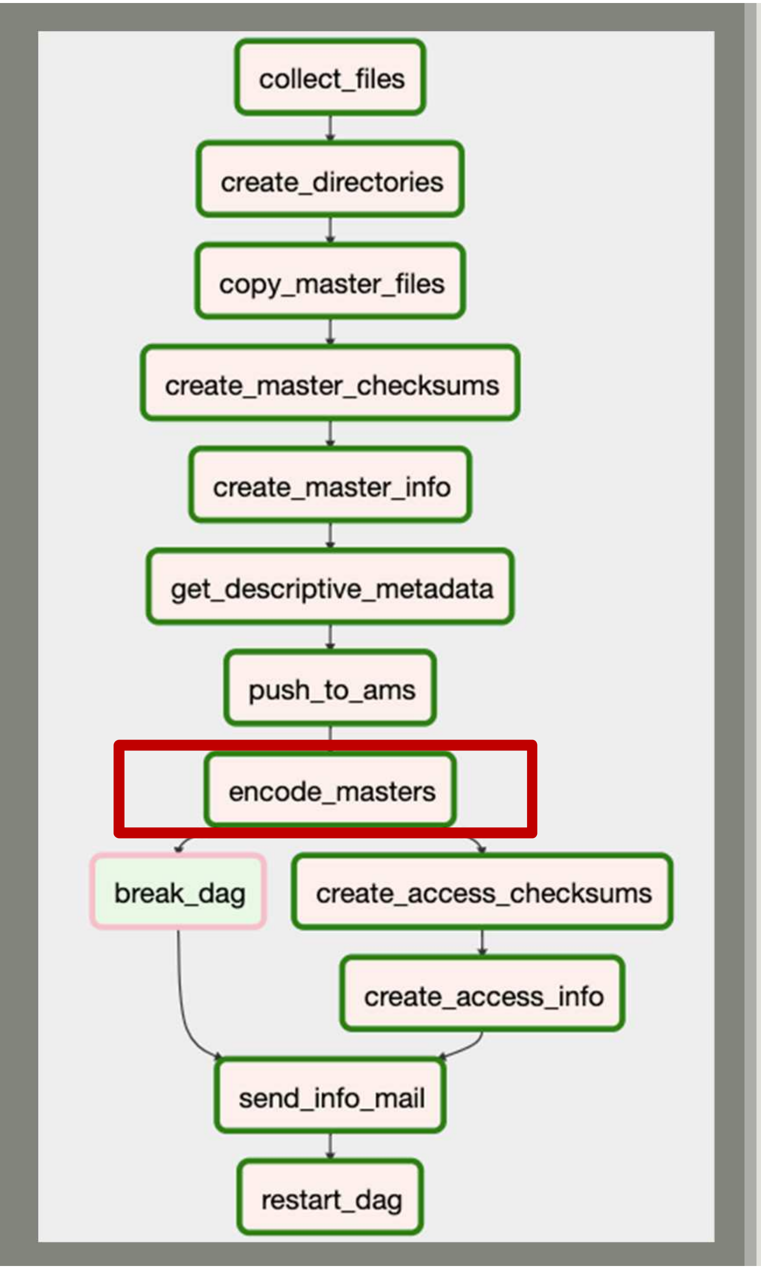
Example:



OSA.Workflow@ceu.edu
Mon 11/4/2019 10:31 PM
Jozsef Bone; Janos Dani

Dear AV team,

It seems that the barcode HU_OSA_00007623 is not registered in the AMS. Please give it a look!
Your sincerely,
AV workflow



Preservation workflow for video files - steps

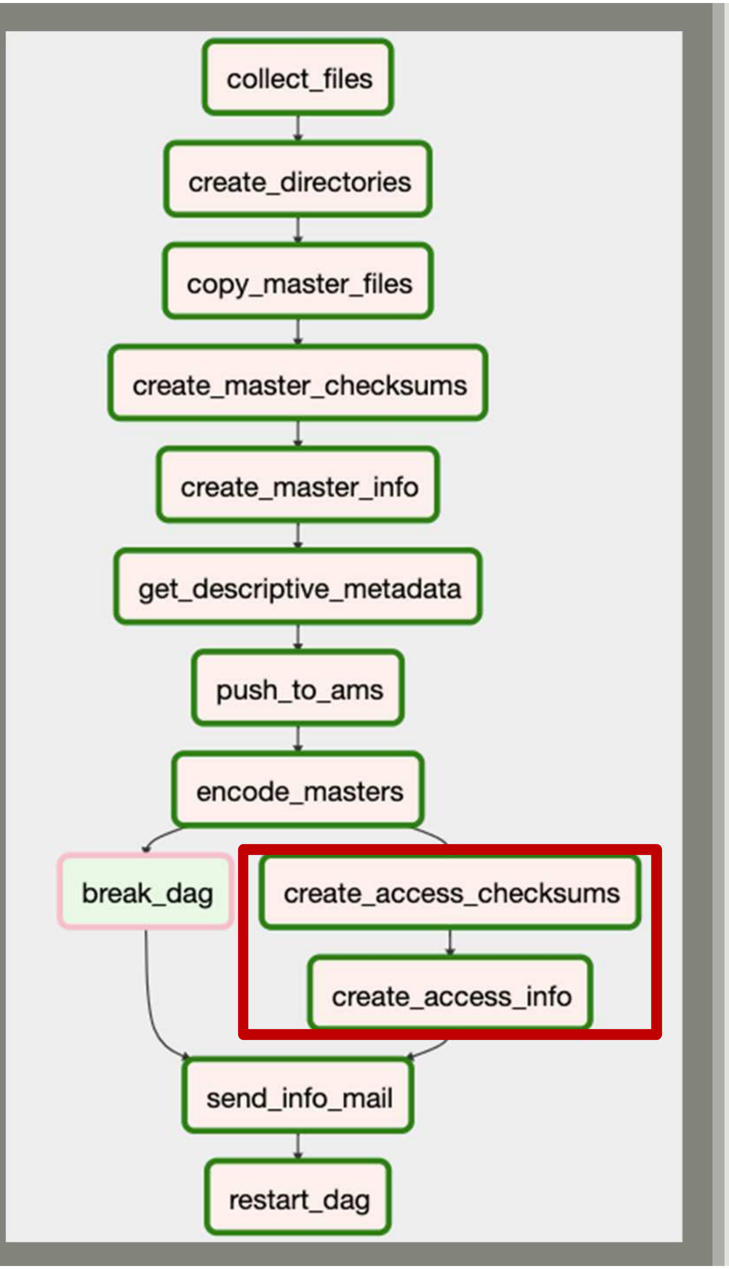
encode_masters

<u>Task:</u>

Create high quality access copy from the master file (h.264 / yuv420p / 7.5M)

Example:

```
HU_OSA_00000011
    Content
        Access
            HU_OSA_00000011.mp4
        Preservation
            HU_OSA_00000011.avi
    Metadata
        Access
        Preservation
            HU_OSA_00000011.md5
            HU_OSA_00000011.sha512
            ...
```



Preservation workflow for video files - steps

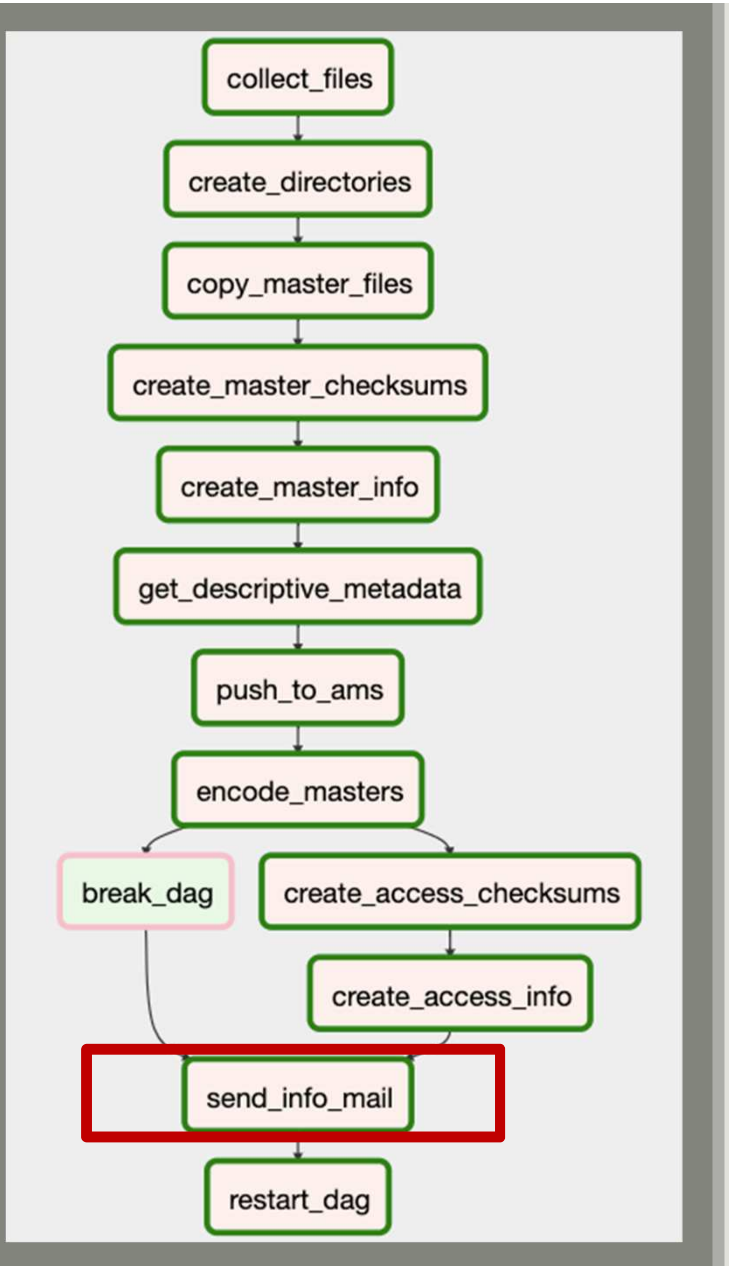
create_access_checksums + create_access_info

Task:

Create md5 and sha512 checksums and technical metadata similar as we did with master files.

Example:

```
HU_OSA_00000011
    Content
        Access
            HU_OSA_00000011.mp4
        Preservation
            HU_OSA_00000011.avi
    Metadata
        Access
            HU_OSA_00000011.md5
            HU_OSA_00000011.sha512
            HU_OSA_00000011_md_tech.json
        Preservation
            ...
```



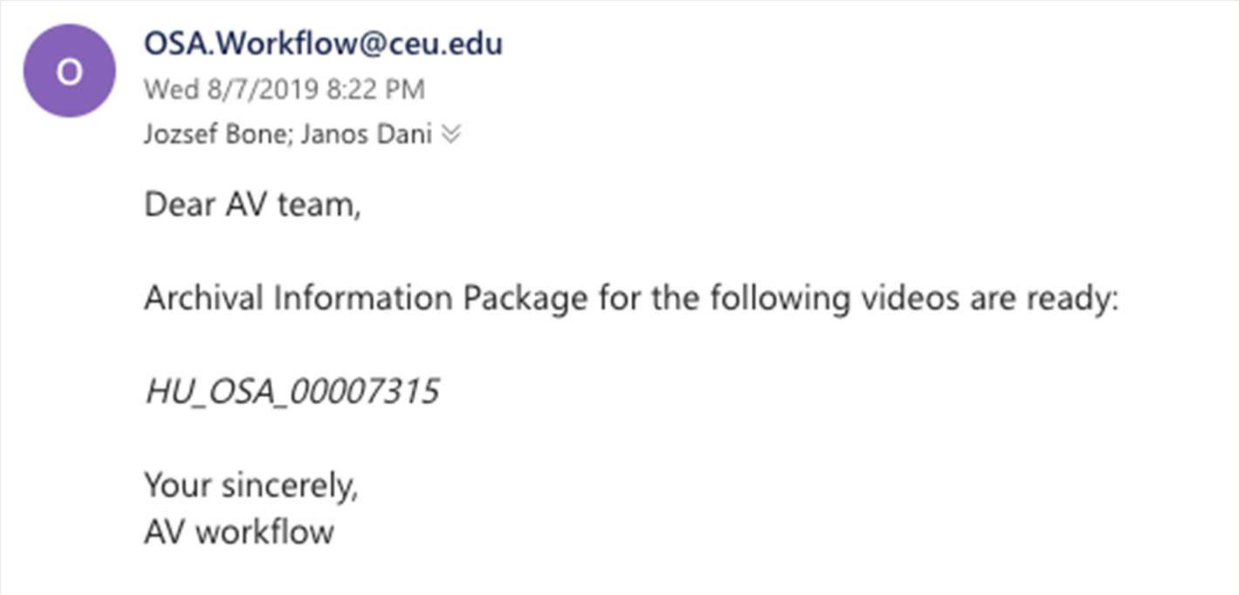
Preservation workflow for video files - steps

send_info_mail

<u>Task:</u>

Send notification email about finishing the workflow.

Example:



O

OSA.Workflow@ceu.edu
Wed 8/7/2019 8:22 PM
Jozsef Bone; Janos Dani

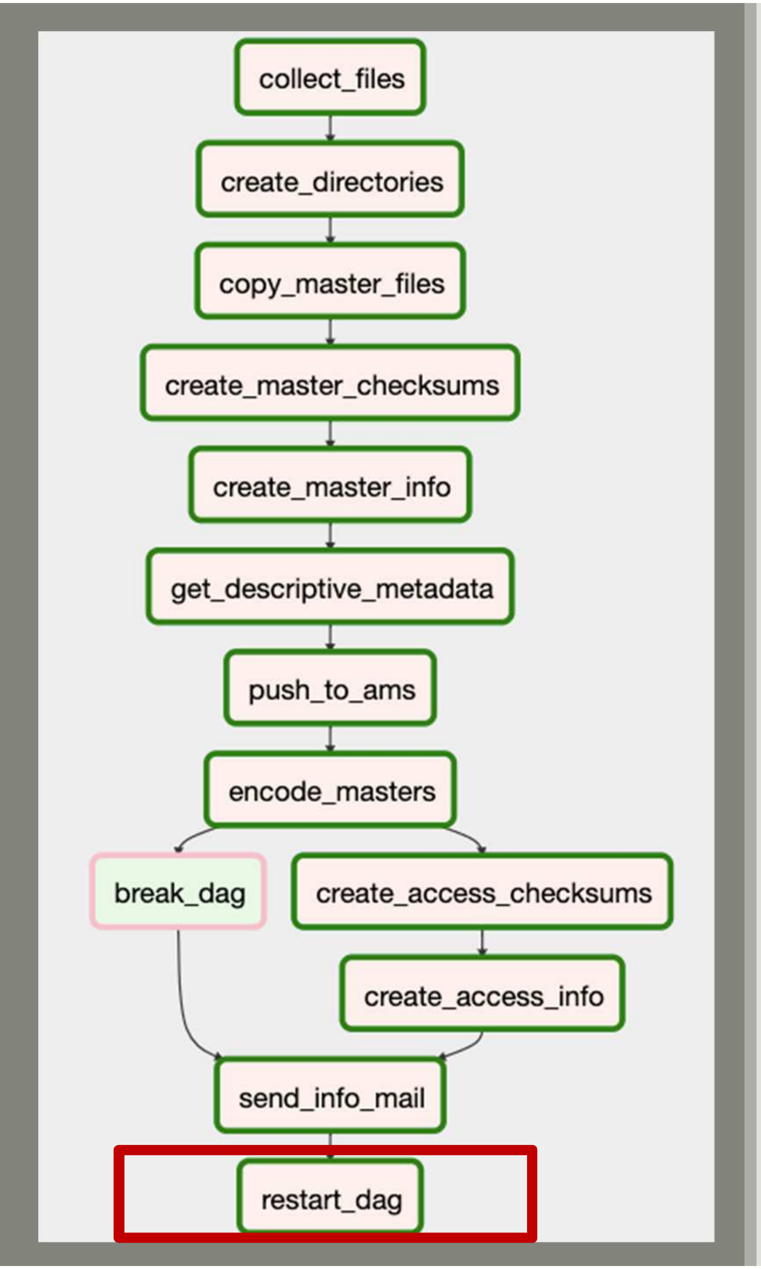
Dear AV team,

Archival Information Package for the following videos are ready:

HU_OSA_00007315

Your sincerely,
AV workflow

Preservation workflow for video files - steps



restart_dag

Task:

Checks if there are master files left in the input directory. If yes, then triggers running the DAG once again, if not exits.

Saving AIP

AIP Structure

```
HU_OSA_00000011
    Content
        Access
            HU_OSA_00000011.mp4
        Preservation
            HU_OSA_00000011.avi
    Metadata
        Access
            HU_OSA_00000011.md5
            HU_OSA_00000011.sha512
            HU_OSA_00000011_md_tech.json
        Preservation
            HU_OSA_00000011.md5
            HU_OSA_00000011.sha512
            HU_OSA_00000011_md_descriptive.json
            HU_OSA_00000011_md_tech.json
```

Storing AIPs

AIPs are written to two simultaneous LTO tapes (currently LTO-7) which will be kept in two separate locations.

A lower quality mp4 file will be made and – depending on copyright – either uploaded to our catalog or to our SharePoint based Research Cloud for internal use.

Errors:

ffmpeg sometimes creates mp4 files with 0 bytes

solution: The encoding task should check the length of the access copy and redo the procedure if the size is 0 bytes.

manual mistakes (missing barcodes in AMS)

solution: Implement a procedure where retriggering certain workflow steps are available with certain signals (like replying an email)

Improvements:

+ splitting DAGs to be able to retrigger certain parts of the workflow.
+ create and add information to PREMIS with every preservation step.
+ include a step to create lower quality mp4 files for web use. Sync them with SharePoint if needed.
+ create other workflows for DVD and audio preservation.
+ create a workflow for automated quality check analysis and feedback for OSA AV staff.

Errors & Improvements

plans for 2020

1200 physical containers were digitized and made available for researchers with this workflow.

So far...

Thank you!

Questions are welcomed.