Archivematica User Forum
Bi-monthly Call: 07/05/2018

Call info:

Meeting agenda and previous minutes can be accessed via Google Drive:
https://drive.google.com/drive/folders/0B4hO2MJSeCa5cmtrclVMOVg0X00

Join from PC, Mac, Linux, iOS or Android:
https://ucla.zoom.us/j/854186191

- Or iPhone one-tap :
  - US: +16699006833,,854186191# or +16465588656,,854186191#
- Or Telephone:
  - Dialc(for higher quality, dial a number based on your current location): US: +1 669 900 6833 or +1 646 558 8656
  - Meeting ID: 854 186 191
  - International numbers available: https://ucla.zoom.us/zoomconference?m=EYLpz4l8KdqWrLdoSAbf5AVRwxXt7OHo

Participants

Please add your name and/or institution below:

- Shira Peltzman, UCLA Library
- Nick Krabbenhoeft, NYPL
- Mira Basara, Cornell University
- Michelle Paolillo, Cornell University
- John Pellman, Columbia University
- Kevin Clair, University of Denver
- Bethany Scott, University of Houston
- Sean Buckner, Texas A&M University
- Jenny Mitcham, University of York
- Bill Kelm, Willamette University
- Andy Weidner, University of Houston
- Courtney Mumma, Texas Digital Library
- Julie Bell, Kansas State University

Convener: Nick Krabbenhoeft

Note Taker: Shira Peltzman

Agenda Items

Please add agenda items below; include your name and/or institution and a suggested length:

I. Assign note-taker (Convener) - *5 min*
   - Note-taker = Shira

II. Welcome - intros for new folks & roll call - *5 min*

III. Topic(s) of the day - *30 min*
   - Desired Features work
     - Check in from John and Shira
       a) John emailed Justin Simpson to mention that it would be nice if there were a GitHub repo to share high-level features requests
       b) Artefactual have been thinking about this as well and have created a new GitHub repo to allow outsiders to create issues etc (https://github.com/archivematica/issues).
       c) They will post to Archivematica Google group next week to explain their vision for this and articulate how the system/GitHub re-organization will work
   - Archivematica Camp
     - Latest camp was 2 weeks ago at Johns Hopkins U.
     - Jenny mentions that there was one in York last April that she helped organize/facilitate. More info in her blog post: http://digital-archiving.blogspot.com/2017/04/archivematica-camp-york-some-thoughts.html
     - Next camp is in Houston: https://wiki.archivematica.org/Community/Camps#November_2018_-_Archivematica_Camp_Texas
     - Perhaps talk more about this in a future call?
   - AICs
     - Who is using them and how?
     - Bethany: at Houston, we are interested in using AICs.
       a) Due to limitations w/ processing capabilities and errors with large AIPs we decided to do one digital object per AIP
       b) Idea to collocate all AIPs from a batch but we have never been able to get AIC functionality to work :(
          (1) Error had to do with creating the AIC METS file
       c) Perhaps in medium-to-long term we could use AICs to do reprocessing

    d) Right now AIC process is entirely manual
- Bill Kelm (Willamette University): not using them here yet
- Jenny Mitcham: we talked about it, but we're not doing it yet
    a) Filling the Digital Preservation Gap report talked about the potential of using AICs to gather together AIPs from the same project (ie, various data that is all related)
    b) Ran into problems with assigning rights info since you can only do this on an AIP-level, and with research data you may have different AIPs with different rights info - we thought AICs may be a way of bringing AIPs from the same researcher together?
- Nick: anyone know when AIC functionality was introduced?
- Courtney Mumma: it started in Canada; there was one institution who wanted this functionality for a research data use case (esp w/ accruals). It was at least 4 years ago (if not longer) University of Alberta funded, here is the documentation https://wiki.archivematica.org/Dataset_preservation
- Shira: thinking about it UCLA, but not at an implementation stage, very similar to Houston process of wanting to collate large amounts of AIPs based on physical objects that are part of a collection
- Sean: are you using the AIC at the collection level, and was this the intent of the AIC?
    a) Shira: yes, and perhaps this is redundant and unnecessary
    b) Bethany: we use Archivematica for processing digitized material and were thinking that if the digitization workflow changes in the future we could use the AIP functionality to help us know which materials need to be reprocessed
- Courtney: whenever someone comes to Artefactual with a features request we (they) try to think of a use case that would apply to other institutions
- Houston "Create AIC METS file" error:

```
Traceback (most recent call last):
  File "/usr/lib/archivematica/MCPClient/clientScripts/createAICMETS.py", line 166, in
<module>
        create_aic_mets(args.aic_uuid, args.aic_dir)
  File "/usr/lib/archivematica/MCPClient/clientScripts/createAICMETS.py", line 151, in
create_aic_mets
        aips = get_aip_info(aic_dir)
  File "/usr/lib/archivematica/MCPClient/clientScripts/createAICMETS.py", line 50, in
get_aip_info
        storage_service.extract_file(aip['uuid'], mets_in_aip, mets_path)
  File "/usr/lib/archivematica/archivematicaCommon/storageService.py", line 355, in
extract_file
        f.write(api.file(uuid).extract_file.get(**params))
  File
"/usr/share/python/archivematica-mcp-client/local/lib/python2.7/site-packages/slumber/__ini
t__.py", line 137, in get
        resp = self._request("GET", params=kwargs)
```

File "/usr/share/python/archivematica-mcp-client/local/lib/python2.7/site-packages/slumber/__init__.py", line 110, in _request
raise exceptions.HttpServerError("Server Error %s: %s" % (resp.status_code, url), response=resp, content=resp.content)
slumber.exceptions.HttpServerError: Server Error 500: https://amtest.lib.uh.edu:8000/api/v2/file/a2b1890f-c089-4d37-b96e-39486f604445/extract_file/

- Handling Large Transfers
  - What makes Archivematica get bogged down more than we want it to?
    a) Nick: NYPL has SIPs that are created w/ 10K files or more in a single package, but bc there's so many of them (even if they aren't large) it can cause the elastic search indexing to time out
       (1) In GitHub repos or Google Group there's a comment about adjusting the elasticsearch timeout, so we have added this to our own instance: once it's up and ready we go to config file and change this from 10 to 100 so that the larger collections won't get bogged down/timed out at this step in the workflow
       (2) Our machines have 8GB RAM
    b) Anyone else who has encountered this?
       (1) John Pellman: we have disabled elasticsearch bc it has been unwieldy w/ large transfers;
       (2) we have also encountered that steps involving FITS will fail bc it consumes memory infinitely, and the file ID microservice will therefore fail. So, we make tweaks to nailgun to cap the memory it can use which has alleviated this problem
       (3) Doing large SIPs is a new frontier; SIPs = 15K items (tifs/jpegs totalling 451 GB)
       (4) We've also had to make tweaks to mySQL settings like log file size setting and (?) → there is another org somewhere who has done this
       (5) We started out w/ a 15 GB RAM instance and eventually upped this to 30 GB RAM (figured out they needed to do this by looking at memory errors and which programs were using the most RAM, and it was always FITS nailgun
       (6) There's a tool that will automatically tell you how much RAM you're using and what's using it
           (a) Tool is sar - it's a operating systems statistics tool that comes with the sysstat package in Red Hat Linux
    c) Jenny: we spent a lot of last year trying to set up an automatic processing workflow for research data; I was trying to push larger and larger things through it

(1) We tried to push through a dataset with 37K files (but they were tiny text files) but we couldn't get Archivematica to work. We never got to the bottom of it but
(2) Justin Simpson thought it may have been the thumbnails that get created, and he filed an issue to fix this; make thumbnails optional. https://github.com/JiscRDSS/archivematica/issues/40
(3) Another issue was removing the standard out and standard errors from the pointer files https://github.com/JiscRDSS/archivematica/issues/39
(4) As we tried to push larger things through Archivematica we did find we ran into problems, and it takes technical expertise to fix the errors that were coming up
(5) Datasets 20 GB in size was our cutoff point

d) Nick: MCP database gets filled up quickly with info that you may not want, especially when you have been processing stuff for a while
e) John: the FPR is in there too, and we have been looking at ways to make the FPR more portable; because of its integration with the MCP the way you may do this is to take the tables and transfer them
f) Nick: default SQL dump tables command and select tables that begin with FPR should do the trick
g) Julie Bell (no mic): We have run 30K files for a 20GB. If we have problems it usually fails on ElasticSearch but we have made a adjustments to Nailgun, ElasticSearch. We run on a R4. large

- How do different institutions handle large transfers?
- Are there specific types of content that take more time?
- What adjustments do you make to default installations to handle large transfers?
- How do we as a group document this information?
    a) Should we post these notes? Communicate this stuff back to Artefactual?
    b) Michelle: this may be a high bar, but it would be great to see this knowledge folded back into a user manual. Knowledge can go out of date as the system changes
        (1) Nick: perhaps we could start communicating w/ Artefactual about this
    c) John: as we encounter bugs, let's open issues on the Archivematica GitHub (not the one I posted earlier but the actual software package - https://github.com/artefactual/archivematica). I have done this previously.

d) Jenny: I've been thinking about how to tie up some of the things we talk about here with the UK Archivematica group, so perhaps I could summarize these discussions and act as a bridge between the two groups from time to time. Other UK users are trying to set up their workflows and perhaps they aren't as far along as the folks on this call but it would be a good idea for future calls. An upcoming topic for the fall call of the UK user group is to experiment with the Appraisal Tab.

- Multiple Archivematica User Groups: UK, Netherlands, Texas, Archivematica User Forum--we should think/talk about how to connect up the convos taking place in each of these

IV. Potential sponsored development work - *10 min*
- [slow summer months, so not currently]
- Remember that if you are thinking of sponsoring something this is a great place to discuss it and find potential partners

V. Conference Roundup - *5 min*
- NDSA
- iPres
  - A few folks have put together a file format signature workshop last fall that will run again at iPRES in September. Not directly Archivematica related but it does feed into PRONOM and therefore the FPR
  - There are 2 papers that Justin Simpson is presenting about checksums and preservation actions registries
  - All workshops and papers can be found here: https://osf.io/u5w3q/wiki/SessionsList/
- OSSArcFlow: https://educopia.org/research/ossarcflow
  - There will be a workshop at iPRES as well, also probably NDSA and other conferences
- Preservation Action Registry - This is a new project to look at how information about preservation actions can be shared by people across platforms. Artefactual are working with Preservica on this one and the work is being supported by Jisc and carried out by the OPF - looks really interesting https://www.artefactual.com/preservation-action-registry-par-research-project-launched/#.Wz5PLNJKiUk (Justin will talk about this more at iPres)

VI. Select next convener & create agenda doc for next meeting (Convener) - *5 min*
- NEXT CONVENER: Bethany Scott!