

Digital Preservation at Oxford and Cambridge

A collaborative research project to evaluate and provide sustainable recommendations for our digital preservation programmes

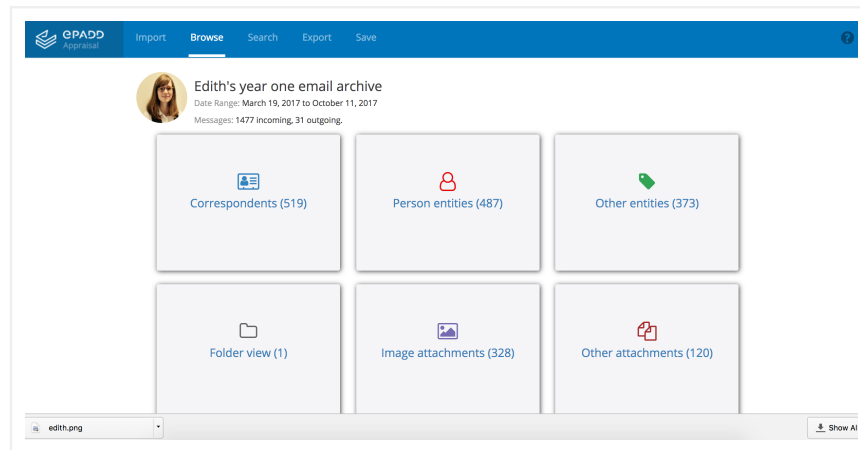
Using ePADD with Josh Schneider

Posted on [23 October, 2017](#) by [ehalvarsson](#)

Edith, Policy and Planning Fellow at Bodleian Libraries, writes about her favourite features in ePADD (an open source software for email archives) and about how the tool aligns with digital preservation workflows.

At iPres a few weeks ago I had the pleasure of attending an [ePadd](#) workshop ran by Josh Schneider from Stanford University Libraries. The workshop was for me one of the major highlights of the conference, as I have been keen to try out ePADD since first hearing about it at [DPC's](#) Email Preservation Day. I wrote [a blog](#) about the event back in July, and have now finally taken the time to review ePADD using my own email archive.

ePADD is primarily for appraisal and delivery, rather than a digital preservation tool. However, as a potential component in ingest workflows to an institutional repository, ensuring that email content retains integrity during processing in ePADD is paramount. The creators behind ePADD are therefore thinking about how to enhance current features to make the tool fit better into digital preservation workflows. I will discuss these features later in the blog, but first I wanted to show some of the capabilities of ePADD. I can definitely recommend having a play with this tool yourself as it is very addictive!



— ePADD: Appraisal module dashboard

Josh, our lovely workshop leader, recommends that new ePADD users go home and try it on their own email collections. As you know your own material fairly well it is a good way of learning about both what ePADD does well and its limits. So I decided to feed in my work emails from the past year into ePADD – and found some interesting trends about my own working patterns.

ePADD consists of four modules, although I will only be showing features from the first two in this blog:

Module 1: Appraisal (Module used by donors for annotation and sensitivity review of emails before delivering them to the archive)

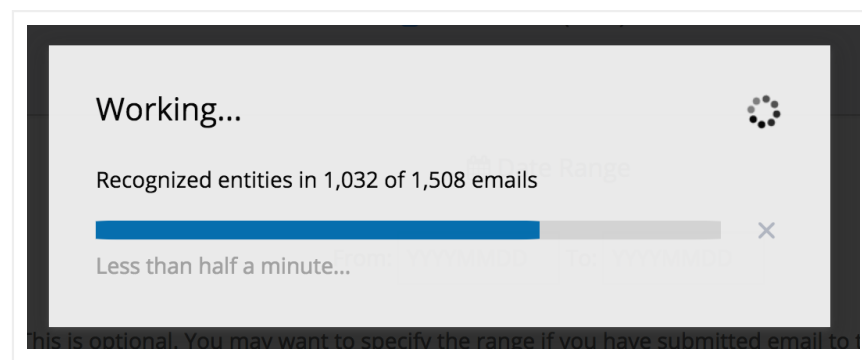
Module 2: Processing (A module with some enhanced appraisal features used by archivist to find additional sensitive information which may have been missed in the first round of appraisal)

Module 3: Discovery (A module which provides users with limited key word searching for entities in the email archive)

Module 4: Delivery (This module provides more enhanced viewing of the content of the email archive – including a gallery for viewing images and other document attachments)

Note that ePADD only support MBOX files, so if you are an Outlook user like myself you will need to first convert from PST to MBOX. After you have created an MBOX file, setting up ePADD is fairly simple and quick. Once the first ePADD module (“Appraisal”) was up and running, processing my 1,500 emails and 450 attachments took around four minutes. This time includes time for natural language processing. ePADD recognises and indexes various

“entities” – including persons, places and events – and presents these in a digestible way.



— ePADD: Appraisal module processing MBOX file

Looking at the entities recognised by ePADD, I was able to see who I have been speaking with/about during the past year. There were some not so surprising figures that popped up (such as my DPOC colleagues James Mooney and Dave Gerrard). However, curiously I seem to also have received a lot of messages about the “black spider” this year (turns out they were emails from the Libraries’ Dungeons and Dragons group).

Type: Person		
Show 10 entries		
Entity	Score	Messages
James Mooney	1	104
Black Spider	1	24
	1	23
	1	17
David Gerrard	1	14
	1	11

— ePADD entity type: Person (some details removed)

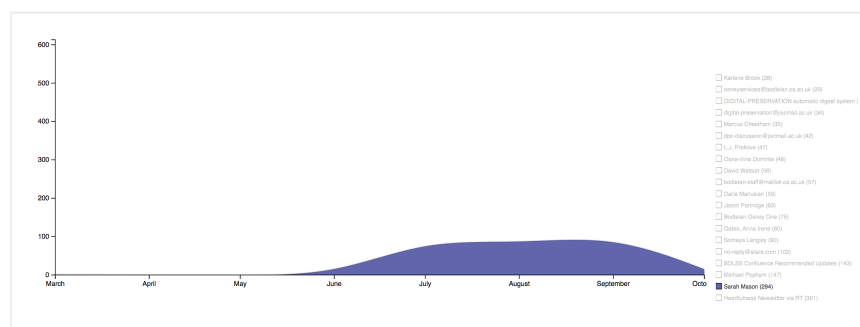
An example of why you need to look deeper at the results of natural language processing was evident when I looked under the “place entities” list in ePADD:

Entity	Score	Messages
San Francisco	1	126
United Kingdom	1	25
Netherlands	1	23
Japan	1	15
St Pancras	1	14
United States	1	12
France	1	11

— ePADD entity type: Place

San Francisco comes highest up on the list of mentioned places in my inbox. I was initially quite surprised by this result. Looking a bit closer, all 126 emails containing a mention of San Francisco turned out to be from “Slack”. [Slack](#) is an instant messaging service used by the DPOC team, which has its headquarters in San Francisco. All email digests from Slack contains the head office address!

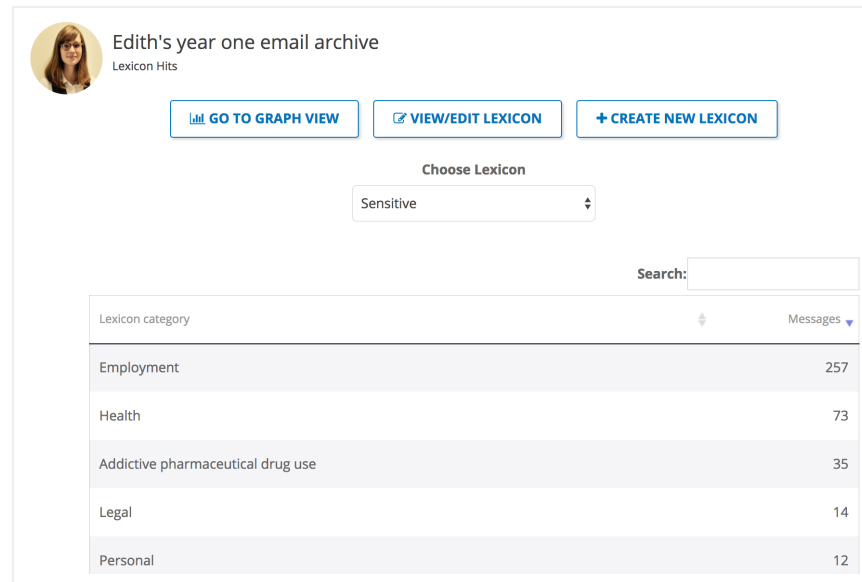
Another one of my favourite things about ePADD is its ability to track frequency of messages between email accounts. Below is a graph showing correspondence between myself and Sarah Mason (outreach and training fellow on the DPOC project). The graph shows that our peak period of emailing each other was during the PASIG conference, which [DPOC hosted in Oxford at the start of September this year](#). It is easy to imagine how this feature could be useful to academics using email archives to research correspondence between particular individuals.



— ePADD displaying correspondence frequency over time between two users

The last feature I wanted to talk about is “sensitivity review” in ePADD. Although I annotate personal data I receive, I thought that the one year mark of the DPOC project would also be a good time

to run a second sensitivity review of my own email archive. Using ePADD's "lexicon hits search" I was able to sift through a number of potentially sensitive emails. See image below for categories identified which cover everything from employment to health. These were all false positives in the end, but it is a feature I believe I will make use of again.



— ePADD processing module: Lexicon hits for sensitive data

So now on to the Digital Preservation bit. There are currently three risks of using ePADD in terms of preservation which stands out to me.

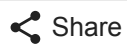
- 1) For practical reasons, MBOX is currently the only email format option supported by ePADD. If MBOX is not the preferred preservation format of an archive it may end up running multiple migrations between email formats resulting in progressive loss of data
- 2) There are no checksums being generated when you download content from an ePADD module in order to copy it onto the next one. This could be an issue as emails are copied multiple times without monitoring of the integrity of the email archive files occurring
- 3) There is currently limited support for assigning multiple identifiers to archives in ePADD. This could potentially become an issue when trying to aggregate email archives from different intuitions. Local identifiers could in this scenario clash and other additional unique identifiers would then also be required

Note however that these concerns are already on the ePADD roadmap, so they are likely to improve or even be solved within the next year.

To watch out for ePADD updates, or just have a play with your own email archive (it is loads of fun!), check out their:

- [Website](#)
- [Git Hub page](#)
- [Twitter](#)
- [Or sign up to their email list](#)

SHARE THIS:



Share

This entry was posted in [born-digital](#), [conference](#), [digital lifecycle](#), [iPres 2017](#), [personal digital archiving](#), [tools](#) by [ehalvarsson](#). Bookmark the [permalink](#) [<http://www.dpoc.ac.uk/2017/10/23/using-epadd-with-josh-schneider/>] .

This site uses Akismet to reduce spam. [Learn how your comment data is processed](#).