

From: [Shulman, Stu](#)
To: [Kasdin, Alexandra](#)
Cc: [Hoy, Mark](#); [Willey, Seth](#); [Sarah Backsen](#); [Cash, Marcia](#)
Subject: Re: Adding files to DiscoverText Project
Date: Saturday, June 18, 2016 5:20:36 AM
Attachments: [image.png](#)

All,

Another minor miracle delivered by Mark! :-)

I have cc'ed Marcia Cash because we are currently talking to DOI about agency-wide perpetual licenses. From what I keep hearing, DiscoverText and our support for technical and methods issues is top shelf. I still believe we could deliver this service government-wide.

Stu

On Fri, Jun 17, 2016 at 2:56 PM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:
It worked!!!! Thanks so much, Mark. This was a HUGE help.

Have a wonderful weekend,
Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Fri, Jun 17, 2016 at 12:45 PM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:
Thanks Mark! I've started the de-duplication. It said it should take 3.33 hours so I'll let you know what it looks like once that's done.

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Fri, Jun 17, 2016 at 12:40 PM, Hoy, Mark <mark@texifter.com> wrote:
No worries -- ok - the new file is there now -- look under the archive named "FWS-Yellowstone_grizzly_delisting_split2" --

- Mark

On Fri, Jun 17, 2016 at 10:46 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:
Sounds great. Thanks Mark! I apologize for all the complications with these files.

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Fri, Jun 17, 2016 at 11:44 AM, Hoy, Mark <mark@texifter.com> wrote:
Hi Alex -

It appears that the Adobe PDF creation / split of the RTF files did not work out quite as it should (it appears to have replaced many "o" characters with "0"s for an unknown reason)... I'm going to try a different approach and extract the text directly from the RTF and use that as the input to DiscoverText... I'll let you know when it's up (should be a lot faster than PDF processing)

On Fri, Jun 17, 2016 at 8:53 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:

Hi Mark,

So unfortunately, it only de-duplicated down to 4,213. In theory, this should de-duplicate down to only one comment (if we can get rid of the signatures and headers). It would be time-consuming for us to sift through 4,213 comments to ensure they are indeed duplicates but we can do it if needed. The method that our colleague used to create the csv files I shared seem to get rid of all the headers. Perhaps if we can load those files in, it would work? Feel free to call me to discuss! (609) 610-2060.

Thanks so much again!

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: (303) 236-4217

On Fri, Jun 17, 2016 at 9:26 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:

One step ahead of you! I am already running the de-duplication. It said it would take about 10 minutes. I will let you know how it goes.

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: (303) 236-4217

On Fri, Jun 17, 2016 at 9:25 AM, Hoy, Mark <mark@texifter.com> wrote:

No worries -- I see however that the new algorithm for extracting the headers and footers got _most_ of the items correct...

I'd suggest running the de-duplication on this, and see if the results are enough that the rest can be ticked off manually.

On Fri, Jun 17, 2016 at 8:23 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:

I see it now! Thank you SO much for your help and patience, Mark. I will let you know if I have any other issues.

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: (303) 236-4217

On Fri, Jun 17, 2016 at 9:20 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:

Her account is sfierce.

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: (303) 236-4217

On Fri, Jun 17, 2016 at 9:20 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:

Yes please! That would be great.

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard

Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Fri, Jun 17, 2016 at 9:19 AM, Hoy, Mark <mark@texifter.com> wrote:
OK. Would you like me to re-assign the archive then to Sarah's account?

On Fri, Jun 17, 2016 at 8:18 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:
Sarah is the administrator of the project but she unfortunately had some schedule conflicts this week. So, we changed Sarah's password so I could use her account and act as the administrator since I was not able to do everything (split dataset, de-deduplicate) from my account (akasdin).

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Fri, Jun 17, 2016 at 9:15 AM, Hoy, Mark <mark@texifter.com> wrote:
Alex -

Looking at the user's list (in the internal admin area for DiscoverText) - I see the account "akasdin" has not been logged into since 6/15 - are you certain that you are on the correct account?

On Fri, Jun 17, 2016 at 8:06 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:
Yes, I am in the archives list for the Yellowstone grizzly project. I tried the shift+F5 and closing my browser. Still no archive, unfortunately!

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Fri, Jun 17, 2016 at 8:58 AM, Hoy, Mark <mark@texifter.com> wrote:
Alex -

Just to double-check - you are on your account, and in archives list for the "Yellowstone grizzly delisting" project, correct?

If you still cannot see it, it might be aggressive caching on the part of your web browser... two more things to try: (1) hold down shift and press F5 at the same time on that page (this will force a cache-busting reload), or (2) log out, completely shut down all copies of your browser, and then try again.

- Mark





On Fri, Jun 17, 2016 at 7:37 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:
I tried logging out and back in and I unfortunately do not see it. Feel free to give my cell a call if that makes it easier to troubleshoot: [\(609\) 610-2060](tel:(609)610-2060)

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Fri, Jun 17, 2016 at 8:35 AM, Hoy, Mark <mark@texifter.com> wrote:
Hi Alex -

I see it -- third one down in the project... if you do not see it still, try logging out and back on, and see if you see it then.

- Mark

Archive Name	Created	Units	Actions
 Center for Biological Diversity form letters	05/23/2016	29,932	
 Endangered Species Coalition form letters	05/23/2016	14,599	
 FWS-Yellowstone_grizzly_delisting_split.zip	06/16/2016	34,067	
 Greater Yellowstone Coalition merged form letters	05/23/2016	2,236	

On Fri, Jun 17, 2016 at 7:31 AM, Kasdin, Alexandra <alexandra_kasdin@fws.gov> wrote:

Hi Mark,

I just wanted to let you know that I don't yet see the archive with the Humane Society form letters. Does your system still show it as uploading?

Thanks!

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Thu, Jun 16, 2016 at 10:41 AM, Kasdin, Alexandra

<alexandra_kasdin@fws.gov> wrote:

Sounds good, Mark! I will let you know if I do not see it tomorrow morning.

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Thu, Jun 16, 2016 at 10:40 AM, Hoy, Mark <mark@texifter.com> wrote:

No worries Alex -

Last I looked this morning, it was still saying an ETA of about 20 hours, so, it may not get fully loaded in and ready until tomorrow morning.

- Mark

On Thu, Jun 16, 2016 at 8:01 AM, Kasdin, Alexandra

<alexandra_kasdin@fws.gov> wrote:

Thank you so much, Mark! I do not see the archive yet. I am sure it is still processing. I will let you know if I have any issues with de-duplication once I see the archive.

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:(303)236-4217)

On Thu, Jun 16, 2016 at 12:25 AM, Hoy, Mark <mark@texifter.com> wrote:

Hi there -

Good news all - tonight, I was able to get the files split (into a total of

34,067 individual comments), and we were able to adjust the algorithm to correctly detect the new header formats in the emails.

I've started the upload into your account Alex, in the project "Yellowstone grizzly delisting", and when it is finished processing, the archive should be named "FWS-Yellowstone_grizzly_delisting_split".

Let us know (once it's there) how things look, and if they look good, you can proceed with deduplication, and near-duplicate clustering if need be.

Thanks!

- Mark

On Wed, Jun 15, 2016 at 8:46 AM, Shulman, Stu <stu@texifter.com> wrote:

An old question from the early days of eRulemaking research. As a matter fact, groups are stubborn about using whatever method they feel like, often throwing us curve balls.

On Wed, Jun 15, 2016 at 10:38 AM, Willey, Seth

<seth_willey@fws.gov> wrote:

It sounds like this is taken care of, but another option we might consider in the future is contacting the submitter and asking if they would send us the comments in an excel file. This is what the FR asked folks to do. In the past, when I've done this folks have been amenable. Food for thought, although it sounds like this issue is mostly addressed.

Seth

Seth L. Willey, *Acting* Chief of Staff
Office of the Regional Director
Mountain-Prairie Region, USFWS
Seth_Willey@fws.gov
[303-236-4257](tel:303-236-4257)

ON DETAIL - May 31st through June 24th, I'll be on a detail as acting Chief of Staff for the Regional Director's Office. Sarah Backsen will be acting for me during this period (sarah_backsen@fws.gov, [303-236-4388](tel:303-236-4388)).

On Wed, Jun 15, 2016 at 8:32 AM, Kasdin, Alexandra

<alexandra_kasdin@fws.gov> wrote:

Hi Mark,

Thank you so much! Please do keep me posted. Again, I am assuming that once we can correctly upload all these files and de-duplicate them, they should compress down to one letter. Fingers crossed!

Alex

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(303\) 236-4217](tel:303-236-4217)

On Wed, Jun 15, 2016 at 12:23 AM, Hoy, Mark

<mark@texifter.com> wrote:

Hi Alex -

We've seen similar requests before and we should be able to help you with this as well. I ran some tests tonight first converting the RTF files to PDF, then splitting the PDFs (using Adobe Acrobat

Pro) into individual PDF files (1 page per file), then zipping and uploading the files into DiscoverText... that got us most of the way there. DiscoverText has built into it algorithms for detecting and converting the headers and footers into metadata, leaving the body text intact as the primary text -- this is functionality we built in a long time ago to help facilitate comment processing for FDMS archives. The algorithms are fairly accurate, however with these documents, it was able to detect the footers accurately, but not able to strip the headers properly.

We will be able to tweak the algorithms to deal with this and we should certainly be able to get these in and processed correctly. Once I have the tweaks in place, I will be able to convert and upload the documents directly into your account. This will probably take a couple of days to get things working correctly, but my hope is to get these all processed and in your account for your use by Friday.

I'll keep you in the loop and let you know how things go. Thanks!

- Mark

On Tue, Jun 14, 2016 at 2:53 PM, Kasdin, Alexandra

<alexandra_kasdin@fws.gov> wrote:

Hi Mark,

I just got off the phone with Stu, who was immensely helpful with some DiscoverText questions I had. We are currently struggling to find the most efficient way to handle importing a form letter into our DiscoverText project. Unfortunately, the Humane Society sent us their form letter submissions in two not-so-ideal formats: PDF and rtf files. They put their signed letters into 69 different files, each with 500 letters (one letter per page). They indicated that the letters were all identical. In theory, DiscoverText should de-duplicate these 34,000+ letters into one individual letter. I trust that these letters are indeed all identical. However, we need the de-duplication report for our record to confirm that they actually are identical and that we considered the comments.

We could not figure out how to import these file types into DiscoverText so it would understand that each page in each of the many documents is one individual comment. Plus, we also needed to figure out how to make DiscoverText ignore the signatures and just focus on the comment content (which was supposedly identical). With the signatures, the comments may look different from each other, even if the text is exactly the same. So, a colleague of ours wrote a script to pull the name of the commenter and the text of the comment into 69 individual Excel files of 500 entries each. I then started to convert these files into csv format and individually upload them into DiscoverText. However, to individually upload each of these 69 files will take another 27 hours with me manually uploading another file every 30 minutes or so. Stu suggested that you might be able to find a more efficient way to input these 34,000+ (supposedly identical) form letter comments into our DiscoverText project. That would be amazingly helpful!

The project is in the account sfierce and is called "Yellowstone grizzly delisting." I started uploading the csv files into the archive called "Humane Society Unedited form letters 1."

Though, you may want to start a new archive and delete that one to avoid duplication of uploading. We aren't super concerned about having the metadata. I am attaching two zip files of these

comments. One zip file contains the rtf files and the other contains the Excel spreadsheets we created from the rtf files.

Feel free to give me a call if any of this doesn't make sense!
Thanks so much in advance for your help. We really appreciate it.

Alex Kasdin
Princeton University SINSI Fellow
Ecological Services, Mountain-Prairie Regional Office
134 Union Boulevard
Lakewood, CO 80228
Phone: [\(609\) 610-2060](tel:6096102060)

--

Dr. Stuart W. Shulman
Founder and CEO, Texifter
LinkedIn: <http://www.linkedin.com/in/stuartwshulman>

--

Dr. Stuart W. Shulman
Founder and CEO, Texifter
LinkedIn: <http://www.linkedin.com/in/stuartwshulman>