

**From:** [Kasdin, Alexandra](#)  
**To:** [stu@texifter.com](mailto:stu@texifter.com); [mark@texifter.com](mailto:mark@texifter.com)  
**Cc:** [Seth Willey](#); [Sarah Backsen](#)  
**Subject:** Adding files to DiscoverText Project  
**Date:** Tuesday, June 14, 2016 3:53:22 PM  
**Attachments:** [Unedited.zip](#)  
[Unedited excel.zip](#)

---

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 files 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