Steps:

1. Get data off the local machine

* Why?
  + Re-archiving lots of small XML files will lock up your processor
  + You might also have storage issues
  + Need the right software to re-archive
* How:
  + Off network file transfer to AWS, Google Drive, etc.

1. Once the data is in a different environment, it needs to be extracted

* This can be done on an EC2 instance, or personal laptop since it is open data. I went with an EC2 micro (free-tier) since my personal laptop didn't have enough storage space. I guess I didn't think to use an external hard drive - whoops. The EC2 micro was really, really slow.

1. Move all XMLs into a single folder

* Data is currently divided into month folders
* To make the final steps easier, all data should be in a single calendar year folder

1. Create a split zip archive with a maximum file size of 75mb

* Requires WinZip or WinRar (I can't remember which I used), maybe 7zip would work too
* 75mb max size is so that we can upload the data to Github for hosting

1. Once everything is re-archived, it needs to be added to the Github repo

* Clone the repo, pull, move files in, commit, then push
* This can be done through command line or through the desktop GUI. It will give an error if you try to do more than one at a time, so upload a single zip folder at a time to get it to work.