ETL Process

In order to do ETL of the database I would first take the latest html file and do a diff with the previous html file that was added to the database. Using diff I would determine what contents are new in our file, and generate a new file with only the new content. After that, I would attempt to run the fixFile.rb script to fix any manual inconsistencies that could hurt the ETL process. Then getSources can obtain the relevant information on the authors.