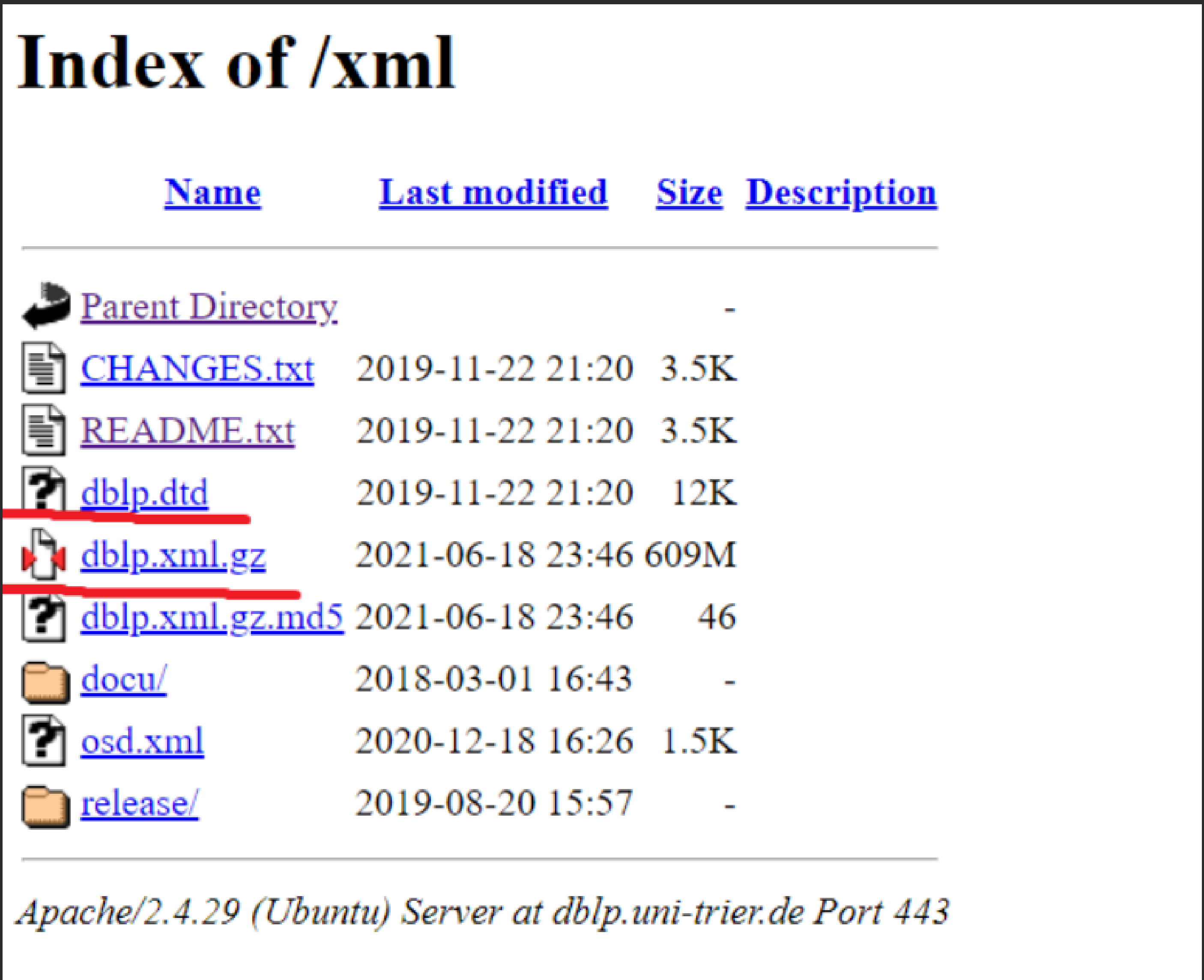
Disclaimer, to update the database one would need around 1TB of storage available and about 30h of computation time.

In order to update the database you must download all 4 data sources and parse them.

1. Create a new folder where all the input files will be stored
2. Go to: https://dblp.uni-trier.de/xml/ and download the dblp.xml.gz and dblp.dtd files

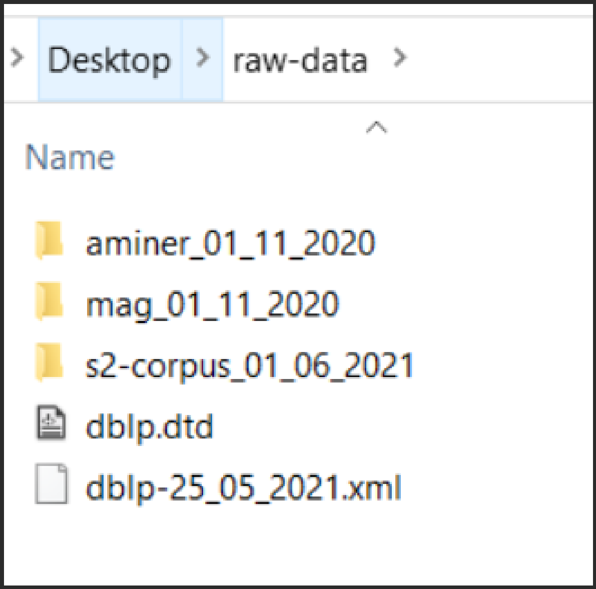


Index of /xml

Name	Last modified	Size	Description
Parent Directory		-	
CHANGES.txt	2019-11-22 21:20	3.5K	
README.txt	2019-11-22 21:20	3.5K	
dblp.dtd	2019-11-22 21:20	12K	
dblp.xml.gz	2021-06-18 23:46	609M	
dblp.xml.gz.md5	2021-06-18 23:46	46	
docu/	2018-03-01 16:43	-	
osd.xml	2020-12-18 16:26	1.5K	
release/	2019-08-20 15:57	-	

Apache/2.4.29 (Ubuntu) Server at dblp.uni-trier.de Port 443

3. Add 3 new folders one for semantic scholar, one for mag and one for aminer data sources, adding the date into the folder name is optional, but it should look like this:



Name

aminer_01_11_2020

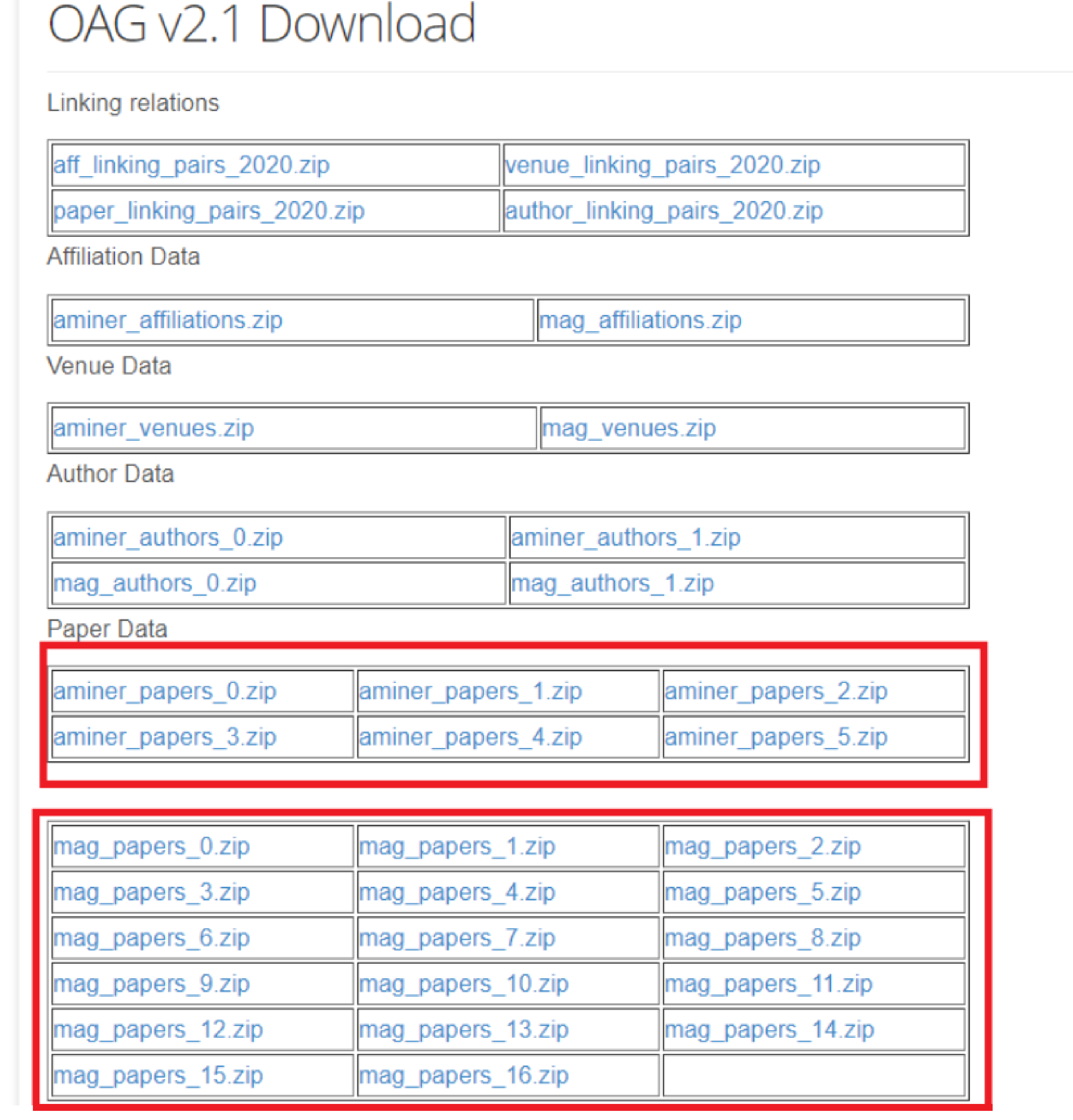
mag_01_11_2020

s2-corpus_01_06_2021

dblp.dtd

dblp-25_05_2021.xml

4. Go to: https://www.aminer.org/open-academic-graph click the latest version available and download all the aminer_papers and mag_papers files, placing them in the aminer and mag directories respectively



OAG v2.1 Download

Linking relations

aff_linking_pairs_2020.zip	venue_linking_pairs_2020.zip
paper_linking_pairs_2020.zip	author_linking_pairs_2020.zip

Affiliation Data

aminer_affiliations.zip	mag_affiliations.zip

Venue Data

aminer_venues.zip	mag_venues.zip

Author Data

aminer_authors_0.zip	aminer_authors_1.zip
mag_authors_0.zip	mag_authors_1.zip

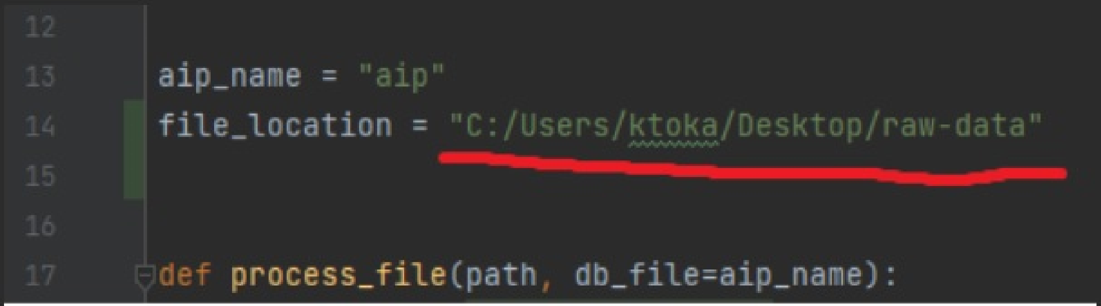
Paper Data

aminer_papers_0.zip	aminer_papers_1.zip	aminer_papers_2.zip
aminer_papers_3.zip	aminer_papers_4.zip	aminer_papers_5.zip

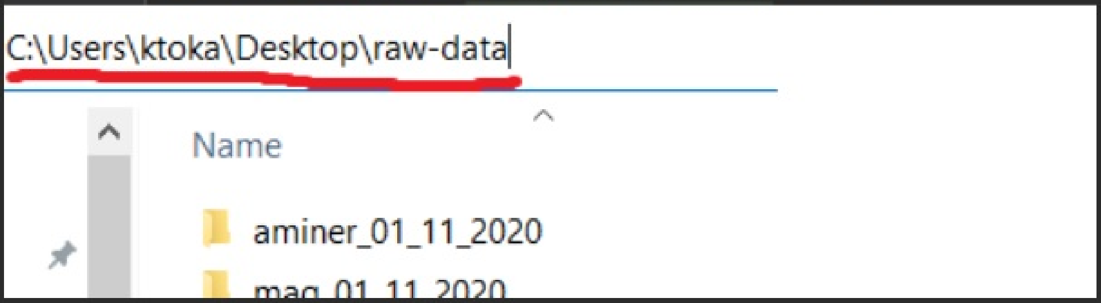
mag_papers_0.zip	mag_papers_1.zip	mag_papers_2.zip
mag_papers_3.zip	mag_papers_4.zip	mag_papers_5.zip
mag_papers_6.zip	mag_papers_7.zip	mag_papers_8.zip
mag_papers_9.zip	mag_papers_10.zip	mag_papers_11.zip
mag_papers_12.zip	mag_papers_13.zip	mag_papers_14.zip
mag_papers_15.zip	mag_papers_16.zip	

5. Go to: http://s2-public-api.prod.s2.allenai.org/corpus/download/ and download all the files by using the wget command on your terminal.
   - Open up the terminal and navigate to the right (s2-corpus directory)
   - Run the "wget https://s3-us-west-2.amazonaws.com/ai2-s2-research-public/open-corpus/2021-06-01/manifest.txt" command to download the manifest text file.
   - Run the "wget -B https://s3-us-west-2.amazonaws.com/ai2-s2-research-public/open-corpus/2021-06-01/ -i manifest.txt" command to download the actual data types. Remember to change the date (here 2021-06-01) to the most recent one available.
6. After downloading all the files, unzip them.

7. After making sure all files are unzipped and stored in the same folder, change line 14 in the renew_data_locally.py, which is located in the parser folder, to the correct path of the folder you downloaded all the files to.



```
12
13      aip_name = "aip"
14      file_location = "C:/Users/ktoka/Desktop/raw-data"
15
16
17    def process_file(path, db_file=aip_name):
```



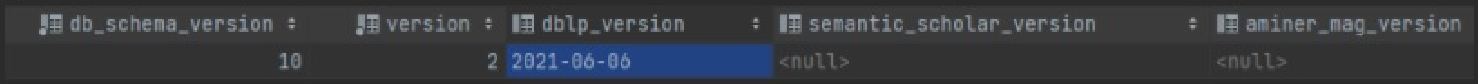
C:\Users\ktoka\Desktop\raw-data

Name

aminer_01_11_2020

mag_01_11_2020

8. Finally, run the renew_data_locally.py file.
9. After re-parsing the whole database, make sure to add the version dates of all the downloaded sources into the database.



db_schema_version	version	dblp_version	semantic_scholar_version	aminer_mag_version
10	2	2021-06-06	<null>	<null>