

iPRES 2024, 19 Sept. - Artiestenfoyer

[Analyzing Large Web Archive Collections Using Parquet Files, Tutorial, 11:00-12:30](#)

Collaborative notes by

- Please add your name here
- Please add your name here
- Please add your name here
- Please add your name here
- Please add your name here
- Please add your name here

Analyzing Large Web Archive Collections Using Parquet Files, Tutorial, 11:00-12:30

Notes:

Definitions:

- WARC: standard for web archives
- CDX: indexes for the WARC having 9 to 11 fields.

Web archive collection is a loose term. A subset depends on the needs of the user.

An index for a web archive collection is a CDXJ file. > storing additional arbitrary metadata in JSON for each record.

Tools for collection analysis:

- Archives unleashed kit
- Archives research computer hub
- CDX summary: <https://github.com/internetarchive/cdx-summary>
- Summarize CDX
- Common Crawl Tools (commoncrawl.org)

Common Crawl:

- Open repository of web crawling data
- They produce a lot of other files than WARC, e.g. Parquet, WAT (extracted metadata), WET (extracted text)

Also possible to access WARC via SQL

End of Term Web Archive: every four year done in US before the elections

- Collaborative web archiving > different partners
- Snapshot of .gov and .mil web every four year
- Saved in WARC, WAT, WET, CDX and META
- <https://eotarchive.org/data/>

Tools used (image taken)

- You can add metadata to a sidecar file with a python tool (warc-metadata-sidecar), e.g. language: <https://github.com/unt-libraries/warc-metadata-sidecar/blob/main/README.md>
- Soft 404 detection: https://github.com/internetarchive/tarb_soft404 a 404 that is from a machine perspective not a 404.
- Also a tool for language identification
- Encoding detection
- In metadata it says “this tool thinks it is this language or this encoding”

Why CDXJ and parquet?

- CDXJ: column based > easier to answer questions that ask for analyzing data
- Parquet is optimized for analytics > easy to run query against

What is Parquet?

- Open source columnar storage file format
- Efficient data compression
- It's a binary file > specialised software is needed
- Easy partitioning, e.g. for every year one
- Can be queries by SQL or as a Dataframe

Slides: <https://github.com/end-of-term/eot-parquet-workshop>