Workshop Tools für Web- und Social-Media-Archivierung: Anleitungen und Links

> ReplayWeb.page: Replay engine
https://replayweb.page/

> Browsertrix-crawler: Crawler
- funktioniert mit docker, kann auch gehostet eingekauft werden
- https://crawler.docs.browsertrix.com/user-guide/

Blogbeiträge:
- https://blogs.bl.uk/webarchive/2024/10/archiving-social-media-with-browsertrix.html: kann auch für Social-Media-Accounts verwendet werden, Login muss bekannt sein
- https://www.onb.ac.at/mehr/blogs/browsertrix-crawling-profile
- https://www.onb.ac.at/mehr/blogs/detail/browser-based-crawling-die-evolution-des-webs
- https://forum.webrecorder.net/t/browsertrix-archiving-of-instagram-posts-showing-up-blank/297/6

Commands:
für Crawl einer Webseite:
docker run -v $PWD/crawls:/crawls/ -it webrecorder/browsertrix-crawler crawl --url https://kastelen.ch --generateWACZ --text --collection test
Crawl eines Instagram-Accounts, man muss eingeloggt sein:
docker run -v $PWD/crawls:/crawls/ -it webrecorder/browsertrix-crawler crawl --url https://www.instagram.com/verein_burgruine_kastelen/ --generateWACZ --text --behaviors --waitUntil load,networkidle2 --collection test
Crawl mit Screencast:
docker run -p 9037:9037 -v $PWD/crawls:/crawls/ webrecorder/browsertrix-crawler crawl --url https://www.instagram.com/verein_burgruine_kastelen/ --screencastPort 9037 --generateWACZ --text --behaviors --waitUntil load,networkidle2 --collection test

> Archiveweb.page: Crawler
- Extension im Chrome-Browser, man muss sich manuell durch eine Webseite klicken (z.B. PDF oder Youtube-Videos werden nur gecrawlt, wenn ich sie aktiv anklicke)
- kann auch als Desktop-App installiert werden: https://github.com/webrecorder/archiveweb.page/releases/download/v0.13.1/ArchiveWeb.page-0.13.1.exe
- Anleitungen: https://archiveweb.page/guide

> Mehr Informationen zu der ganzen Webrecorder-Familie:
- https://webrecorder.net/
- https://github.com/webrecorder

> Nicht am Workshop gezeigt:

- Auto-Archiver, von Bellingcat
- https://github.com/bellingcat/auto-archiver; funktioniert mit docker
- Link auf mein Google-Sheet: https://docs.google.com/spreadsheets/d/19UrmoKNK18O0YwdCHbzmIqO7yJ50QfryMtuquFqKMHY/edit?pli=1&gid=0#gid=0

- Archive It, Testzugang stand leider noch nicht bereit; Infos zum Testaccount: https://support.archive-it.org/hc/en-us/articles/360046164091-About-trial-access-to-Archive-It

- Brozzler, https://github.com/internetarchive/brozzler: distributed browser-based web crawler, basiert auf Python, schwierig auszuprobieren, da verschiedene Dienste; nicht auf docker vorhanden; https://github.com/internetarchive/brozzler; Archive It verwendet u.a. Brozzler

- Twint für automatisierte Twitter-Archivierung: https://github.com/twintproject/twint, braucht Python, wird nicht mehr weiterentwickelt; https://archive20.hypotheses.org/10031