Open Access Database Server (OADS) URL Crawl

Dr. Jian Wu, Spencer Peloquin

The purpose of this experiment was to crawl and classify the OADS research repository to uncover which disciplines were losing research data to URL link rot for future remediation.

URL link rot is an ongoing problem in digital academia, with numerous studies and research being lost due to changes in access and indexing. In order to maintain the permanence of critical research in the online domain, it is necessary to discover which studies are most impacted and the reasons why this occurs.

The first step in this process was to crawl the OADS database for nine months from April to February to track how many links were lost each month. This was done using a python bot that would scrape OADS every month on the tenth day. The next task was to classify the URLs of one month (April) to provide a snapshot of which disciplines were most impacted by the URL rot. This was done using a python classification framework that would convert the JSON data of April into a CSV file that could be analyzed for trends in data loss.

The crawler found that the number of URLs available (Error Code 2xx) decreased linearly from April to October, briefly resurged in November, and resumed its previous trend of linear decay from December to February. This was expected, as the number of valid URLs is expected to decrease linearly as the rate of access loss is likely higher than the rate of new papers published. November's anomaly is within the margin of error and may be due to a configuration issue rather than any surge in publication. Finally, category 0 (biology) is the most commonly lost research category because of the sheer size of papers published, rather than any specific problem with published papers in this discipline.

The findings show that high volume publications have a much larger rate of URL link rot. This decay greatly exceeds the publication speed. With this knowledge, it can be shown that much of the decay comes from standard network changes rather than deliberate or unintentional delisting. The high likelihood of category 0 being lost shows that larger publication cohorts are hit the hardest by URL link rot. Thus, future steps should be taken to back up and re-upload data while continuing to regularly publish. This also underscores the necessity of third-party backups in research beyond the major publishers. Perhaps a future study can target correspondence between archival (Wayback?) backup and extant publications.