Preserving Open Access Datasets and Software for Sustainable Computational Reproducibility

Jian Wu (ODU CS), Sawood Alam (Internet Archive), William A. Ingram (VT Libraries), Edward A. Fox (VT CS)

1 Introduction

The *Old Dominion University (ODU)*, in collaboration with the Internet Archive and the Virginia Polytechnic Institute & State University (Virginia Tech), proposes to lead a *3-year Applied Research* project aiming at preserving endangered Open Access Datasets and Software (OADS), defined as publicly and freely available digital datasets and software packages used for producing research results reported in academic documents. Our proposal is aligned with Objectives 3.1–3.3 under *Goal 3 (Improve the ability of libraries and archives to provide broad access to and use of information and collections)* of the National Leadership Grants for Libraries Program.

Our preliminary study based on 16 AI papers (Ajayi et al. 2023) indicated that 6 out of 16 papers (38%) contain accessible data *and* executable codes, while 4 out of these 6 papers (67%) reported reproducible results, which highlighted the importance of the *availability* and *accessibility* of OADS in reproducing computational results in academic literature. However, a significant fraction of URLs link to OADS that claimed to be accessible but no longer are accessible (Salsabil et al. 2022; Escamilla et al. 2023). This situation undermines the FAIR (findable, accessible, interoperable, reusable) Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016) and has become a hurdle in verifying published results. To mitigate this situation, it is necessary to automatically and reliably identify and analyze OADS-URLs, defined as URLs linking to OADS, from scholarly documents, represented by academic papers and electronic theses and dissertations (ETDs) in this project. Another challenge is how to predict and preserve endangered OADS. We will address these challenges and aid a wide spectrum of stakeholders in multiple disciplines in achieving sustainable reproducibility of their scholarly results. The intended project results include software packages, manually and automatically extracted data, and enhanced archival and digital library services.

2 Project Justification

Recently, concerns of reproducibility have been raised in multiple academic disciplines such as the Social and Behavioral Sciences (Camerer et al. 2018), Biomedical and Life Sciences (Gannot et al. 2017), and Computer and Information Sciences (Collberg & Proebsting 2016; Pimentel et al. 2019; Dacrema et al. 2021). Datasets and software packages (Reinecke et al. 2022) are critical resources to reproduce computational results in disciplines including but not limited to artificial intelligence (Hutson 2018), and domains requiring data analysis (Seibold et al. 2021). Collberg and Proebsting have found that a large fraction of works in Computer Science were not reproducible because the code and/or data were not available. In part due to advocacy for open science (Collberg & Proebsting 2016), an increasing number of authors choose to share datasets and software publicly. Further, the White House Office of Science and Technology Policy (OSTP) issued guidance in 2022 to make federally funded research freely available without delay. However, our recent research indicates that a substantial fraction of OADS is not archived or are only archived in one archival repository, posing a danger for the academic and industrial communities to reproduce or replicate research results (Escamilla et al. 2023). Therefore, it is urgent to identify these resources and preserve them for sustainable reproducibility.

Automatically identifying OADS from scholarly documents at scale is non-trivial. Our recent work indicated that it was possible to train a hybrid classifier to distinguish OADS-URLs and non-OADS-URLs based on the linguistic features of the URLs' context sentences (Salsabil et al. 2022). However, more challenges still exist on how to obtain fine-grained classifications, such as distinguishing URLs linking to only datasets or software, and how to prioritize important and endangered OADS to be archived. Our proposed research will fill these gaps. The scope of this project is well aligned with all three objectives of Goal 3. For Objective 3.1, our project will create an effective ranking model to sustain a robust online environment for reproducibility. For Objective 3.2, our project will develop new state-of-the-art approaches in natural language processing to support preservation and access to scanned and born-digital content (OADS). For Objective 3.3, our project will design and develop discoverable and accessible archival services to meet the expectations of a wide spectrum of academic and industrial researchers.

3 Project Work Plan

The goal is to develop, report about, and solve the foundational problems related to the value, status, trends, and preservability of OADS for publicly available academic papers and ETDs, and to enable and ensure the advancement toward sustainable computational reproducibility. The data scope of our proposal will cover about 8.1 million academic papers with full text in the Semantic Scholar Open

Access Corpus (S2ORC; Lo et al. 2020) and over 500K ETDs we collected from university library repositories in the United States (Uddin et al. 2021). The time scope will cover from the 1990s until 2020 (for S2ORC) and from the 1990s until 2023 (for ETDs).

To this end, we propose several research questions (RQs) and key tasks (Table 1). To better support the proposed tasks, and to evaluate and guide the overall progress, we will form an *advisory board* consisting of web scientists and librarians from information science departments, university libraries, and the Library of Congress. We will set a higher priority to recruit graduate research assistants from underrepresented students in *Information Science*. ODU is a minority-serving institute, with more than 25% minority students.

Table 1: An outline of research questions, tasks, and results of the proposed project.

Yr	Research Questions, Tasks, and Results
1	**RQ:** How to automatically and accurately identify OADS-URLs from academic documents at scale?
	Task 1 (exploratory): OADS-URL Extraction and Classification. We propose a context-aware few-shot learning model based on GPT and compare it against supervised, pre-trained, and transfer-learning models. **Result:** A software package that automatically and accurately extracts and classifies URLs from scholarly documents (i.e., OADS-URLs) into five classes depending on the OADS resource types and providers.
	Task 2 (implementation): Identifying Open Access Scholarly OADS. We will apply the result in Task 1 to about 8.1M academic papers and 500K ETDs, and check accessibility of OADS pointed to by these OADS-URLs from the live Web and a snapshot (i.e., mementos; Sompel et al. 2009) from the Internet Archive. **Result:** A corpus of 1M+ automatically classified OADS in multiple disciplines within a wide time range. The dataset that links OADS-URLs to archived OADS will help researchers find datasets and software in a broad set of domains, ensuring one-stop-shopping coverage of most US OADS. The analysis will show the relative coverage status of OADS in papers vs. ETDs (where the nation's university research is reported in great detail).
1-2	**RQ:** What is the distribution of accessible OADS across different disciplines and how fast are these resources disappearing from the Web?
	Task 3: (exploratory): Analyzing OADS-URLs across disciplines and time. We will break down the computational reproducibility into three dimensions: *availability* (whether OADS-URLs are present in the paper), *discoverability* (whether OADS are live or archived), and *accessibility* (whether the OADS can be found on the OADS-URL pages). Using temporal data from the Internet Archive, we will study how the three components change over time, how they vary across disciplines, and how fast OADS disappear. **Result:** The distribution of OADS resources across multiple academic disciplines and across decades of time span, providing quantitative guidance of web archiving.
2-3	**RQ:** How to predict which OADS should be archived, and how to rank archiving priorities.
	Task 4 (exploratory): Building a Predictive Model for Endangered OADS. Using historical data from the Internet Archive, we will train state-of-the-art machine learning models to predict endangered OADS that may soon disappear, as well as models to rank archiving priorities. **Result:** Effective and efficient predictive models to identify and rank endangered OADS based on their importance, and the risk of disappearance.
3	**RQ:** How do we preserve them and make them accessible using web archives and digital libraries?
	Task 5 (exploratory and deployment): Integrating Archived OADS into Digital Libraries. We will investigate the technical and policy challenges of archiving and preserving OADS to align with the Desirable Characteristics of Data Repositories for Federally Funded Research. We will link OADS-URLs to CiteSeerX and *Fatcat Wiki* (hosted by Internet Archive), to serve a broad spectrum of scholars in various domains. **Result:** Technical requirements and policies. Permanent OADS-URLs linked to CiteSeerX and *Fatcat Wiki*.

4 Budget Summary

The estimated total budget is $575,967 over the 3-year period. As the lead institution, ODU requests $420,461 for direct costs and $155,506 for indirect costs. Salary and fridges are being applied to PI Wu and two graduate student research assistants in *information sciences* ($191,776, plus $36,285 tuition). The subcontract to Internet Archive ($100,000) will fund co-PI Alam's salary, fringes, travel, and infrastructure costs. The subcontract to Virginia Tech ($75,000) will fund senior personnel Fox and Ingram to aid research on a weekly basis. We allocate a budget of $5,400 to compensate 6 advisory board members, who will participate in two meetings each year.