# Narrative

The *Old Dominion University (ODU)*, in collaboration with the Internet Archive (IA) and the Virginia Polytechnic Institute & State University (Virginia Tech), proposes a *3-year Applied Research* project for preserving endangered Open Access Datasets and Software (OADS), i.e., publicly and freely available digital datasets and software packages used for reproducing research results reported in scholarly works. We focus on scholarly papers (journal articles and conference proceedings) and electronic theses and dissertations (ETDs) in multiple disciplines. 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 (NLG) for Libraries Program, and Objective 3.2 under IMLS agency-level *Goal 3 (Advance Collections Stewardship and Access).*

## 1 Statement of National Need

Recently, concerns of reproducibility have been raised in multiple academic disciplines such as the Social and Behavioral Sciences [4], Biomedical and Life Sciences [16], and Computer and Information Sciences [8, 36, 15]. Datasets and software packages [37] are crucial resources to many research domains requiring data analysis [41]. Collberg and Proebsting 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 [8], an increasing number of authors choose to share datasets and software publicly. Further, the White House Office of Science and Technology Policy [23] 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, posing a barrier for the academic and industrial communities to reproduce or replicate research outcomes [12]. Therefore, it is urgent to identify and preserve endangered OADS resources for sustainable reproducibility.

Our preliminary study on a focused AI research topic [2] indicated that only 6 out of 16 papers (38%) contain accessible data *and* executable codes, while only 4 out of the 6 papers (67%) reported reproducible results. This highlights the importance of OADS in reproducing computational results in academic literature. Yet, a significant fraction of URLs link to OADS that claimed to be, but are no longer, accessible [40, 12]. This situation undermines the FAIR (findable, accessible, interoperable, reusable) Guiding Principles for scientific data management and stewardship [47] and has become a hurdle in verifying published results. To mitigate this situation, it is necessary to automatically and reliably identify *OADS-URLs, i.e., URLs linking to OADS*, from *scholarly works*, represented by *scholarly papers* and *electronic theses and dissertations (ETDs)*, and then to effectively predict and preserve endangered OADS before they disappear. Automatically identifying OADS from scholarly documents at scale is challenging. Our recent work trained a hybrid classifier to distinguish OADS-URLs and non-OADS-URLs based on the latent linguistic features of the URLs' context sentences [40]. However, further challenges exist on how to obtain fine-grained classifications, distinguishing the provenance (author-provided vs. third-party) and specific types (datasets vs. software), and how to prioritize endangered OADS to be archived. We will address these challenges and aid a wide spectrum of stakeholders including but not limited to educators, students, researchers, web archivists, and librarians. Intended project results include software packages, manually labeled and automatically extracted data, and enhanced digital library and archival services.

The scope of this project is well aligned with all objectives of Goal 3 of the NLG Libraries Program. For Objective 3.1, our project will create effective ranking models and investigate the feasibility of creating and sustaining a robust online reproducibility environment. For Objective 3.2, we will develop new approaches to support preservation and access to scanned and born-digital content. For Objective 3.3, we will design and develop discoverable and accessible archival services to support a wide spectrum of academic and industrial users.

**The goal of this project is to develop, report about, and solve foundational problems related to the value, status, trends, and preservability of OADS for publicly available scholarly papers and ETDs, and to enable and ensure progress toward sustainable computational reproducibility.** To this end, we will focus on building machine learning models and datasets that encompass three key aspects of OADS, namely,

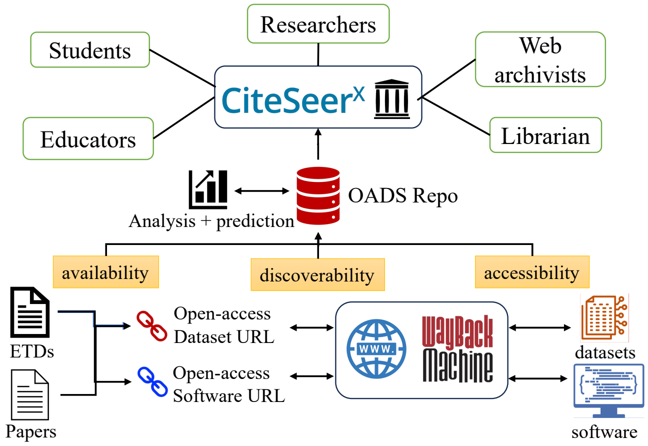

**Figure 1:** The schematic overview of the project proposed, including main components and stakeholders.

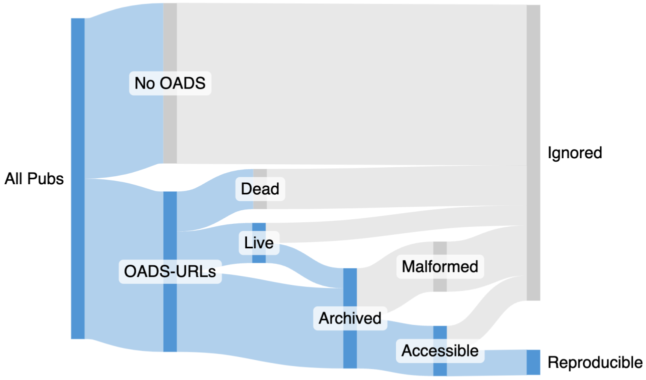

**Figure 2:** Illustration of a gradually decaying prevalence of availability, discoverability, accessibility, and reproducibility (thicknesses are not proportional to yet-to-be-assessed reality).

*availability* (whether OADS-URLs appear in scholarly works), *discoverability* (whether OADS-URLs are alive on the web or in the archive), and *accessibility* (whether OADS are accessible through OADS-URLs). Figure 2 illustrates our main contributions, including validating the three aspects, the OADS Repo dataset to be built, analytical and prediction models, and dissemination platforms (CiteSeerX and Internet Archive).

***Research Scope:*** We propose tasks that generate answers to these questions. These tasks would yield deliverables in the form of tools, models, and standards. Beyond the scope of this project are: (1) establishing operational infrastructure, and (2) building a new archive or repository. Instead, we will focus on building a framework and testing it on at least one existing repository. The OADS framework aims to produce results in existing (e.g., RDF Triplets) or newly defined standard formats that can be imported by various repositories and archives.

## 2 Project Design

This project is partially designed atop a prior IMLS proposal on mining ETDs, during which we collected a corpus consisting of more than 500K ETDs from institutional repositories hosted by U.S. university libraries. Our four research areas address the four research questions (Table 1).

### 2.1 Datasets

One contribution of our project is an automatically constructed database called OADS Repo, built by automatically applying machine learning models on scholarly papers and ETDs. To train and evaluate such models, a labeled dataset must be constructed first as the ground truth.

**Labeled Dataset:** The labeled dataset will consist of at least 1,000 OADS-URLs, their context sentences, and associated metadata from arXiv papers and ETDs. A *context sentence* contains the sentence that includes an OADS-URL. Two examples are shown below, in which the magenta color URL links to an open-access dataset and the blue color URL links to an open-access software package.

> The dataset is publicly available and searchable online at http://www.nrel.gov/lci/database/.
> All of the presented structural figures were produced using pymol (http://pymol.sourceforge.net).

The metadata contains basic fields of the paper the URL belongs to, such as the publication year and dsciplines (i.e., subject categories). The labeled dataset characterizes availability, observability, and accessibility. . Each URL will be manually labeled into 5 *availability* categories, based on the types and providers of resources, namely, *author-provided dataset, author-provided software, third-party dataset, third-party software, and general URLs.*

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

| |
|---|
| **RQ1:** How to automatically and accurately identify OADS-URLs from academic documents at scale? |
| **Task 1 (exploratory): OADS-URL Extraction and Classification.** We propose Large Language Model (LLM)-based and supervised machine learning (ML) models that extract and classify URLs from scholarly works into five classes depending on resource types and providers. |
| **Task 2 (implementation): Building the Largest OADS-URL Corpus.** We will apply the result in Task 1 on scholarly papers and ETDs and build OADS Repo, the largest repository containing 10 million OADS-URLs and metadata, covering multiple disciplines and a wide time range. |
| **RQ2:** What are the distributions of accessible OADS across disciplines and how fast do they disappear? |
| **Task 3: (exploratory): Analyses across disciplinary and chronological dimensions.** We will study how availability, discoverability, and accessibility change over disciplines and time, providing quantitative guidance of web archiving.<br>**Task 4 (exploratory): How Fast OADS Disappear.** We will analyze the longevity distribution of OADS-URLs and OADS, and build regression models to predict their lifetime. |
| **RQ3:** How to predict which OADS should be archived, and how to rank archiving priorities? |
| **Task 5 (exploratory): Building predictive models for endangered OADS.** Using historical data from the Internet Archive, we will train ML models to predict endangered OADS that may soon disappear. We will investigate the feasibility of preserving them using archival services. |
| **RQ4:** How do we preserve them and make them accessible using web archives and digital libraries? |
| **Task 6 (exploratory and deployment): CiteSeerX as a Testbed to Disseminate OADS-URLs.** We will link OADS-URLs to CiteSeerX [7] and *Fatcat Wiki*, to serve a broad spectrum of users in various domains. |

For training and analysis purposes, we will balance the sample size across these categories. The *discoverability* category can be labeled by manually verifying whether the OADS-URL is alive or archived at the time it is verified. The *accessibility* category can be labeled by manually verifying whether the OADS actually exists by following the OADS-URL at the time it is verified.

The multiclass labeling schema will extend the binary classification schema in our preliminary work [40]. This schema will allow us to put more insight into nuances of the *provenance* and *function* of the OADS resource. In addition to labels, the labeled data also contains properties of OADS-URLs at various levels. These properties will be used as features to train machine learning models such as the models to automatically classify an arbitrary URL into one of the five categories above, given its context sentence (see Section 2.2.1). The dataset will be independently labeled by two graduate students. A consensus rate (Fleiss' $\kappa$) will be calculated as an evaluation metric of data quality. The full schema of the labeled data can be found in Appendix A.

**Full Dataset:** This full dataset called OADS Repo is estimated to contain about 10 million OADS-URLs, with context sentences and associated metadata, extracted from about 17 million (before deduplication) scholarly works obtained from 4 sources: Semantic Scholar Open Research Corpus (S2ORC; [30]), PubMed Open-Access Subset (PMOAS;[10]), arXiv [19], and ETDs [45]. This dataset will be automatically extracted using the state-of-the-art URL extractor and classifier. We will automatically collect the metadata that comes with the data or by querying the OpenAlex API service [34]. The full dataset will be used for correlation and trend analysis (Section 2.2.2), training prediction models for endangered URLs (Section 2.2.3), and linking to digital libraries (Section 2.2.4). The details of methods to build the dataset will be included in Task 2. The description of each dataset is in Appendix B.

**MI Subset:** Recent research reveals that underrepresented groups produce higher rates of scientific novelty but their contributions were devalued and discounted [22]. To facilitate studying the OADS preservation status of underrepresented groups, we will build a special subset for minority institutions (MI), including Historically Black Colleges and Universities (HBCUs) and Hispanic Serving Institutes (HSIs). Our ETD collection includes at least 3K ETDs from HBCUs. Author affiliation of scholarly papers will help us identify papers with leading authors from HBCUs or HSIs.

## 2.2 Research Areas

Our research covers four major areas, each focusing on answering an RQ designated in Table 1.

### 2.2.1 Research Area 1: OADS-URL Identification

The first research area is to explore how to effectively and efficiently identify OADS-URLs from scholarly works. This will be carried out in two tasks. Task 1 will develop the backbone ML models for OADS-URL extraction and classification. Task 2 will apply this model to about 17 million scholarly papers and ETDs (before removing duplidates) to build the full dataset. We first briefly review recent advancements in related work.

**Related work**

*URL extraction:* Not all PDF files are created equal when it comes to embedding hyperlinks in them, depending on the tool used to create the PDF file. Some PDFs may contain anchor-texts or URLs that are clickable, which is usually facilitated by adding hyperlinks in the *Annotation* layer of the PDF pages. These links are easier to extract, but can be malformed at times. In some PDFs, the links are present as plain text and the long ones are broken on multiple lines with or without hyphenation, which would require sophisticated pattern matching to extract them reliably [3, 21].

*URL classification:* Existing work on URL classification aims at classifying URLs into malicious vs. normal URLs, or by content topics. Heuristic methods and learning-based methods have been explored. Most methods are based on features extracted from URLs. Zhao et al. [51] proposed a multitask learning method to classify URL citation context by embedding the citation context by BERT [11], a language model that converts text to dense vectors. Roman et al. compared word embedding methods in the task of citation context classification using a dataset consisting of 10 million citation contexts and achieved an F1-score of 81% using an unsupervised classifier [38]. In another paper, the authors proposed new features (section title and footnote text) as well as BERT, to classify URLs in scholarly papers by their functions [44].

*Computational reproducibility:* Computational reproducibility has been studied in several recent papers. One studied the URLs linking to datasets, focusing on papers produced by ACM SIGMOD and PVLDB [35]. The authors used a simple keyword-based method to search for links to source materials. If the link was found active, they considered the resource to be available. In another study, tf-idf and cosine similarity were used[17]. Färber et al. analyzed the quality and usage of GitHub code repositories using MAG [13] and found a strong bias towards specific computer science areas (e.g., ML) and publication venues. In another work, authors studied 1.4 million Jupyter notebooks from GitHub [36] and found that only 24.11% of notebooks executed without errors and only 4.03% produced the same results. URLs used in these two studies were limited to GitHub links and therefore papers containing these URLs were published mostly after 2010 when GitHub was launched.

**Research activities**

***Task 1: Effective and Efficient OADS-URL Extraction and Classification.*** The goal of this task is to develop the backbone ML models to automatically extract and classify OADS-URLs effectively and efficiently from scholarly works. We will first experiment and compare open-source software packages to convert PDF files to text such as PDFMiner [42], PDFPlumber [43], PyPDF2 [14], and PyPDFium2 [24], based on the number of characters, words, and URLs in the output. We then segment free text into sentences using NLTK's sentence tokenizer. However, many URLs do not directly appear together with the sentences. Instead, they appear in footnotes or are cited in bibliography sections. To overcome this challenge, we will build a module to restore URLs back to sentence context using heuristic approaches, in which a series of regular expressions will be explored to identify the footnote or citation marks and link URLs back to the positions in the sentence contexts they belong to. The *left* and *right* sentences will also be extracted if they exist. This preprocessing pipeline will prepare the text to build the labeled data. To ensure data quality, the context sentences of URLs will be manually verified before human labeling. Our previous binary hybrid classifier that relies on embedding the context sentence achieved an F1-score of 92%. However, when applied to newly labeled data consisting of 5 classes, the model failed to achieve the same level of performance because it could not capture semantic nuances expressed in the sentence. We propose two methods to tackle this problem by improving the data representation and the classification models.

*Method 1* is the *co-training with expanded context.* Co-training is a semi-supervised learning technique that trains two classifiers based on two different views of data [27]. In our case, we will train one classifier based on URLs and the other classifier based on expanded context, including the *left* and *right* sentence of the target context sentence. Increasing the context span enriches semanic features, which can potentially improve classification accuracy. Co-training has been shown to be effective in sentiment classification of MOOC forum posts [5].

*Method 2* will harness the power of LLMs. In particular, we will explore the role of *reasoning* using GPT3.5 for URL context classification. We will evaluate state-of-the-art prompting methods such as Chain-of-Thought (CoT) [46] in which users first present some examples each containing a series of steps of thoughts before reaching a conclusion and then request GPT to mimic the reasoning process. CoT has boosted the performance of many arithmetic and commonsense tasks.

We will compare the efficiency, efficacy, and scalability between co-training and LLM-based methods. The deliverable of this task is a pipeline that extracts and classifies OADS-URLs from scholarly papers and ETDs.

***Task 2: Building the Largest OADS-URL Corpus.*** We will build the largest OADS-URL corpus based on about 17M scholarly works from S2ORC, arXiv, PMOAS, and ETDs. This can be carried out in three steps. The first is to remove duplicate paper records that appear in more than one source dataset. S2ORC contains external IDs that were used to identify papers cross-listed in arXiv and PubMed. The second step is to extract and classify URLs. We will take two measures to ensure data quality for further analysis. The first is preferentially adopting full text in XML or LaTeX formats. Specifically for arXiv papers, we will leverage the cleansed text in the unarXiv dataset [39]. For scanned documents we will compare at least two open-source OCR tools (e.g., Tesseract, docTR [31]) by the number of output tokens and URLs. If no one tool exhibits a clear advantage, we will use an ensemble method to choose output with better quality. The second measure is automatic error correction because when converting PDF to text, most text extractors will make grammatical and spelling errors. We will leverage state-of-the-art grammar and spell checkers to automatically correct these errors such as Enchant [29].

Next, we will automatically measure the availability, discoverability, and accessibility of OADS. The availability is measured by the appearance of OADS-URLs using the extractor and classifiers developed in Task 1. The *discoverability* is measured by querying the URL against the Web or the Internet Archive and checking the response code. The Internet Archive has a similar service that is used by their Reference Explorer[1] tool to assess the health of references in Wikipedia pages. The *accessibility* is whether the OADS can be found on the OADS-URL pages. Directly validating the integrity of the resources is a multi-step process and usually requires domain knowledge. Here, we will use heuristic and learning-based methods to predict the existence of data files or software packages. We will levrerage the OADS-URL labels and properties collected in the labeled dataset to build ML classifiers and/or LLM-based prompts and infer the accessibility. Finally, the metadata at various levels (see Appendix A) will be retrieved from data sources and integrated into the full dataset. The dataset will facilitate the subsequent analyses, predictions, and dissemination.

[1] https://archive.org/services/context/iare/

### 2.2.2 Research Area 2: Chronological and Disciplinary Analyses

This RA will answer RQ2 by analyzing OADS-URLs across time and disciplines. Specifically, we will conduct analyses in the three dimensions (availability, discoverability, and accessibility) respectively and estimate how fast OADS-URLs and OADS disappear.

**Related works**

*Link Rot Studies:* Klein et al. reported that one in five scholarly articles suffer from reference rot [28]. In a follow-up study Jones et al. reported that three out of four URL references lead to changed content when dereferenced [26]. Some authors proposed a standard called Robust Links, to introduce HTML element attributes that systematically encode archived version of URLs (i.e., URI-Ms) and the target time of archived version [25]. The IA has deployed a bot called InternetArchiveBot that analyzes pages from various language editions of Wikipedia to catalog links and check their HTTP status code periodically. Once they identify a broken link for a significant amount of time (i.e., beyond the doubt of transient errors), they replace the link with a corresponding archived link [20]. However, it is not guaranteed that a good version of every external link from Wikipedia pages is present in web archives, so the Internet Archive now proactively archives every link added to any Wikipedia page and advertised via their EventStream API[2] to future-proof them. They have rescued more than 20 million broken links (and counting) by February 2024, across various Wikipedia language editions[3]. Establishing a similar pipeline for proactively preserving links in scholarly publications would be desirable, but it comes with unique challenges such as a diverse set of sources to collect the signals from and difficulties in identifying potential OADS links.

*OADS-URL Analysis:* Existing studies [40] found that the availability of OADS-URLs in ETDs has been constantly increasing over time since 2000. The authors also found that the availability exhibited a skewed distribution across multiple disciplines. The discoverability was studied in [12], which found that the 4 mainstream Git Hosting Platforms (GHPs; GitHub, GitLab, SourceForge, and Bitbucket) account for only 33% OADS-URLs; non-GHP OADS-URLs are distributed across almost 50,000 unique hostnames. Existing studies are either limited to ETDs or arXiv papers [12]. The analysis did not distinguish between dataset and software, author-provided and third-party-provided. The *accessibility* was not studied.

**Research Activities**

***Task 3: Analysis in Two Dimensions:*** In this task, we aim to study the chronological trends of availability, discoverability, and accessibility of OADS using the data collected in RA1. We have three objectives. The first is to verify the trends found in ETDs [40] using a much larger and more diverse sample with fine granularity. Specifically, we will study availability represented as the normalized fractions of scholarly works (papers or ETDs) that contain OADS-URLs in a particular year, and delineate it across multiple disciplines. A similar study will be performed for *discoverability* and *accessibility*. The results will reveal the status of the three dimensions over time. The dependency on disciplines will indicate whether disciplines should be used as a feature to predict endangered OADS (Section 2.2.4). In particular, we will make comparisons between MIs and non-MIs. The disparity can shed light on the resource distribution status between these two types of institutions.

***Task 4: How Fast OADS-URLs Disappear:*** In this task, we will focus on the discoverability and accessibility of OADS, aiming at estimating how fast OADS-URLs and OADS disappear. Take OADS-URLs as an example. Here, we assume each OADS-URL has a single life, starting from $t_0$. For author-provided OADS-URLs, $t_0 = t_p$, in which $t_p$ is the year when a scholarly work was published. We will adopt a retrospective view by looking back at the percentage of OADS-URLs $p(\Delta t, d)$ that disappeared for scholarly works published $\Delta t = t - t_0$ years ago for a certain discipline $d$, in which $t$ is the current year. The decay curve $p'$ vs. $\Delta t$, in which $p' = 1 - p(\Delta t, d)$ will allow us to calculate the half decay time of OADS-URLs. Previous studies found that this half decay time was about

[2] https://wikitech.wikimedia.org/wiki/Event_Platform/EventStreams

[3] https://tarb.sawood-dev.us.archive.org/

8 years using biomedical articles [50]. Our study will extend the investigation using a much higher volume of scholarly works, across diverse disciplines. The results will reveal the nuances of trends of disappearing OADS-URLs, and provide quantitative insights on the differential urgency to preserve endangered OADS in various disciplines. We will also compare the decay curves for MI and non-MI and investigate the reason that may cause the difference. This will shed light on the preservation policy adjustment to balance different types of academic institutions.

### 2.2.3 Research Area 3: Predictive Model for Endangered OADS

Our third research area builds on findings in RA1 and RA2 with an investigation of ML models to predict endangered OADS-URLs, by estimating the probability $P(l)$ an OADS-URL $l$ disappears at the time of prediction. $P(l)$ will be used for determining priorities to archive OADS the OADS-URLs link to.

**Related works**

The link rot prediction is closely related to web crawl planning. Recently, ML methods such as Support Vector Machines [52] and Random Forests [1] were employed for this task. Acuna et al. also used a Tobit model, which is a linear regression model, to estimate the longevity of resources shared in scientific publications [1].

**Research Activities**

***Task 5: Building a Predictive Model for Endangered OADS:*** The goal of this task is to build an effective model to predict which OADS are likely to disappear so the corresponding OADS-URLs should have higher priority to be preserved. We propose a supervised classification model train on *all* OADS-URLs whose *discoverability* is measured. The input features include various properties of OADS-URLs and the time they have been alive, and the predicted label is determined by the probability an OADS-URL is alive after $\Delta t$. We will investigate linear classifiers, such as logistic regression, and non-linear classifiers, such as a feedforward neural network with a non-linear activation.

Based on the predictions, we will build a software framework that periodically predicts endangered OADS-URLs and automatically crawls and preserves OADS. This framework will allow us to investigate the technical feasibility of deploying a long-term archival service to preserve OADS, focusing on space capacity, bandwidth requirements, and time scale. The experiments will be conducted at IA.

### 2.2.4 Research Area 4: Integration into Digital Libraries

In this RA, we will investigate the technical challenges of preserving OADS to align with the Desirable Characteristics of Data Repositories for Federally Funded Research. We will develop software to support linking OADS-URLs to digital libraries, using CiteSeerX [18] as a testbed.

**Research Activities**

***Task 6: CiteSeerX as a Testbed to Disseminate OADS-URLs:*** CiteSeerX is a digital library search engine hosting over 15 million scholarly works, including papers and ETDs [18]. CiteSeerX provides searching, metadata browsing, and PDF downloading services. The system has been accessed by 100,000-180,000 users daily [49], making it an ideal testbed to validate if the OADS-URL data extracted will benefit stakeholders. The system was refactored in the past three years to make the service more sustainable. The new infrastructure consolidates the data storage and search functions into a single cluster using Elasticsearch. PI Wu codirects the CiteSeerX project with Dr. C. Lee Giles at the Pennsylvania State University (letter attached).

In this task, we will link OADS Repo data obtained in Section 2.1 to CiteSeerX and display classified OADS-URLs on the paper summary pages. The new CiteSeerX repository ingests papers from the Semantic Scholar, arXiv, and PMOAS. Therefore, most papers used for building OADS Repo should have counterparts in CiteSeerX. The challenge is that there is not an ID as a key to map papers from CiteSeerX to OADS Repo. Our preliminary work on near-duplicate detection shows that the Locality Sensitive Hashing (LSH) was a promising method for this task, achieving an F1-score of 0.846 when tested on the CiteSeerX data and is very scalable [48]. LSH has been used as an efficient method to resolve near-duplicate news articles by Google [9].

The reason we did not directly extract OADS-URLs from CiteSeerX papers is that CiteSeerX contains a significant fraction of non-academic documents due to the imperfection of its academic filters. Our approach will potentially save time and computational resources for data processing. We will evaluate the impact of OADS-URLs using the following metrics: clickthrough rate over time, top-clicked OADS-URLs, and the number of feedback messages. The feedback messages allow users to correct metadata errors and thus reflect the attention of end users. We will also analyze the subject categories of papers associated with OADS-URLs to understand the disciplinary distribution of end users.

### 2.3 Evaluation

Evaluations are performed for each task towards a deliverable. ***Task 1*** will develop the backbone extraction and classification pipeline. The evaluation metrics will include the precision, recall, and F1-scores of OADS-URLs extracted from the benchmark dataset. We will compare the co-training and LLM-based approaches against other baseline approaches, and choose the best one after trading-off between performance and scalability. The evaluation metrics will be calculated for each category. Our goal is to reach at least 90% for each category.

***Task 2*** will generate the OADS-URL corpus. The evaluation will focus on the data quality. We will use a stratified method to sample 500 papers from S2ORC, PMOAS, and arXiv, and 500 ETDs, respectively (so in total 2000 scholarly works) and manually verify the measurements of availability, observability, and accessibility, using standard classification metrics including precision, recall, and F1-scores.

***Tasks 3 and 4*** will be chronological and disciplinary analyses. The evaluations will focus on the significance of dependent variables on independent variables. To that end, we will calculate uncertainty at each data point. For example, the uncertainty of the number of OADS-URLs in a certain year will be characterized by propagating errors caused by the classification model and by the number of papers in that year. The decay curves of OADS-URLs and OADS can be evaluated by quantifying the uncertainty of curve fitting using root-mean-square (RMS) errors.

***Task 5*** will deliver the predictive model, which will be evaluated using a held-out set from the historical snapshots from the Wayback Machine. For the probability model, we will threshold the output probability into binary values and evaluate it using standard metrics including precision, recall, and F1-scores.

***Task 6*** will be evaluated through user traffic, including clickthrough rates and the number of feedback messages. We will use traffic from minority institutions, including HBCUs and HSIs, by tracking their IP addresses.

### 2.4 Project Management

#### 2.4.1 Project Team

This project will proceed at ODU, IA, and Virginia Tech. Responsibility for management, research, and dissemination will be shared between PI Wu and Co-PIs Alam, Fox, and Ingram. The project will hire two graduate research assistants at ODU. The graduate assistants will undertake data collection, labeling, programming, analysis, development, and deployment. They also will assist with publishing and other dissemination activities.

**Dr. Wu** will serve as the PI at ODU. Wu works with the Web Science Digital Library (WS-DL) group to research the mining of scholarly big data, including millions of academic papers. Wu has been the tech leader of CiteSeerX since 2013 and published 80+ peer-reviewed papers on document classification, information extraction, and other topics related to scholarly big data. Dr. Wu will lead Tasks 1, 3, 5, and 6. He will supervise the two graduate research assistants.

**Dr. Alam** will serve as a Co-PI for this project. At IA, Alam will extract metadata and features from mementos of OADS-URLs from Wayback Machines and conduct proof-of-concept experiments on OADS preservation policies. He will co-advise the two graduate students at ODU. Dr. Alam will lead Tasks 2 and 4.

**Mr. Ingram** is Associate Dean and Executive Director for IT in the University Libraries at Virginia Tech. He will be senior personnel. He will attend weekly meetings and provide advice on technical and policy aspects about digital documents, data acquisition, and how to maximize the impact on librarians.

**Dr. Fox** is a full professor at Virginia Tech. He is Director of the Digital Library Research Laboratory, and Executive Director of the Networked Digital Library of Theses and Dissertations. Dr. Fox will be senior personnel. He will attend weekly meetings to provide advice on algorithms, system architecture, web archiving, and how to maximize the impact on students and researchers.

### 2.4.2 Advisory Board

To aid us in evaluation and performance management, we have assembled an advisory board to meet with the project team regularly, evaluate our progress, and keep us on track. The tentative advisory board includes six members: *Dr. Trevor Owens*, a librarian, researcher, policy maker, and educator. Dr. Owens used to work at the Library of Congress and is now at the American Institute of Physics, holding a position at the intersection of libraries, archives, museums, publishing, and higher education. *Karen Vaughan* is the head of scholarly communication and publishing at Old Dominion University Libraries. She has been a professional librarian for 39 years primarily in research support, teaching, digital collections and repositories, and scholarly communication. *Dr. J. Stephen Downie* is a Professor and the Associate Dean for Research of the School of Information Sciences at the University of Illinois, Urbana-Champaign. He is Co-Director of the HathiTrust Research Center. Dr. Downie's research interests focus on the design and evaluation of information retrieval systems. *Dr. Suzie Allard* is the CCI Associate Dean for Research, Director of the Research and Innovation Center, Chancellor's Professor, and the CCI Board of Visitors Professor at the School of Information Sciences at the University of Tennessee, Knoxville. Dr. Allard's research interests include Knowledge Creation and Data Management. *Dr. Martin Klein* is a Scientist at the Los Alamos National Laboratory Research Library. Dr. Klein focuses on research and development efforts in the realm of web archiving, scholarly communication, digital system interoperability, and data management. *James R. Jacobs* is the Federal Government Information Librarian at Stanford University Library. He is a member of the Government Documents Roundtable (GODORT) of the American Library Association and a former chair of GODORT's Government Information Technology Committee (GITCO) and Publications Committee.

## 2.5 Project Dissemination and Sustainability

We plan to release all software created by this project as open-source, hosted in PI Wu's Lab's GitHub repository (see Data Management Plan for details). We will share our research findings through appropriate conferences and journals, such as the Joint Conference on Digital Libraries (JCDL), International Conference on Theory and Practice of Digital Libraries (TPDL), International Journal on Digital Libraries (IJDL), the Innovative Applications of Artificial Intelligence Conference (IAAI), the Association for Information Science and Technology Journal (ASIS&T), the Coalition for Networked Information (CNI), the Open Repositories Conference, and the International Symposium on ETDs. We will also introduce data and software to institutional libraries by hosting workshops and tutorials co-located with conferences. PI Wu and Co-PI Fox have hosted tutorials at the JCDL, where Fox has co-chaired a series of Web Archiving and Digital Library (WADL) workshops. PI Wu will incorporate ML, text mining, information retrieval, and other components into graduate and undergraduate courses, such as on Deep Learning. Other research output generated by this project—including blog posts, preprints, news reports, presentations, derived data sets, software, and other digital products—will be preserved and made available through the Old Dominion University Digital Commons.

## 3 National Impact

The proposed research project will have a national impact for its first investigation of computational reproducibility from the perspectives of *availability*, *discoverability*, and *accessibility* of OADS and linked resources, by mining the *full text* of scholarly papers and ETDs. Our research will impact the information science research community: (1) to build novel methods to automatically and accurately extract and classify OADS-URLs from scholarly works using ML models; (2) to quantitatively reveal the urgent need to preserve OADS resources and improve the situation of computational reproducibility in various disciplines; (3) to provide novel models using archived data to rank the priority to preserve OADS and identify key factors that make an OADS-URL last longer; (4) to enhance digital library services by providing classified OADS-URLs and link them with scholarly papers and ETDs to a widely used digital library search engine; (5) to aid Librarians to enrich the library's digital collection, promote data reuse and sharing.

The project also has broader impacts beyond the digital library research community. Because OADS-URLs ubiquitously exist in scholarly papers and ETDs, the improved computational reproducibility will potentially benefit a spectrum of users from *various disciplines* to access OADS and instruct the best practices to preserve OADS for sustainable access. This can substantially reduce the time researchers spend on reproducing and replicating published studies. The predictive models are potentially important to understand the resource disparity between researchers at MI and non-MI institutes and provide evidence-based policy guidance to balance the distribution of web archiving resources, which is especially critical for underrepresented researchers who have limited resources to preserve data and software. By incorporating ETD collections, our research will potentially advance the integration between research and education of graduate-degree seeking students. Further, incorporating the research into undergraduate and graduate teaching will disseminate interdisciplinary knowledge of information sciences, library sciences, web sciences, and artificial intelligence, and potentially strengthen the research capacity of higher education.

## 4 Assumptions and Potential Risks

**Assumptions:** This research project assumes that URLs and their context sentences can be extracted from scholarly papers and ETDs. We observed a fraction of scholarly papers and ETDs were scanned from photocopies and the text extracted using open-source OCR or commercial OCR engines yielded little text and/or gibberish characters. These exceptions are usually not predictable but can be easily detected, e.g., using extractor error messages or by output file sizes. We will try to employ the best text extractors available by comparing them using standard metrics. Accordingly, we assume the papers and ETDs in our data are written in English and published in PDF format, which reflects the majority of highly impactful publications.

**Potential risks:** Although we focus on open-access scholarly papers and ETDs, it is extremely challenging to obtain an unbiased set that represents all the open-access scholarly papers and ETDs. To mitigate this selection bias, we will report the analysis results based on normalized statistics, and quantify the uncertainties using statistical methods.

| **Project Year 1** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Activities and Milestones** | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul |
| **Administrative work** | | | | | | | | | | | | |
| Launch project | X | | | | | | | | | | | |
| Plan plenary advisory meetings | X | | | | | | | | | | | |
| Advisory Board Meetings | X | | | | | | X | | | | | |
| Hire graduate research assistants | X | X | | | | | | | | | | |
| Create and maintain a project website | | X | X | X | X | X | X | X | X | X | X | X |
| Evaluate project progress | X | X | X | X | X | X | X | X | X | X | X | X |
| Review data management plan and digital products plan | X | | | X | | | X | | | X | | |
| Annual reporting | | | | | | | | | | | X | |
| | | | | | | | | | | | | |
| **Research Area 1**: OADS URL Identification | | | | | | | | | | | | |
| Task 1: Effective and Efficient OADS-URL Extraction and Classification. (lead by Wu) | X | X | X | X | X | X | | | | | | |
| Task 2: Building the Largest OADS-URL Corpus. (lead by Alam) | | | | X | X | X | X | X | X | X | X | X |
| | | | | | | | | | | | | |
| **Research Area 2**: Chronological and Disciplinary Distributions (lead by Wu) | | | | | | | | | | | | |
| Task 3: Analysis in Two Dimensions (lead by Wu) | | | | | | | | | | | X | X |
| | | | | | | | | | | | | |
| *All investigators will collaborate and make contributions. | | | | | | | | | | | | |

| **Project Year 2** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Activities and Milestones** | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul |
| **Administrative work** | | | | | | | | | | | | |
| Advisory Board Meetings | X | | | | | | X | | | | | |
| Hire graduate research assistants | X | X | | | | | | | | | | |
| Create and maintain a project website | X | X | X | X | X | X | X | X | X | X | X | X |
| Evaluate project progress | X | X | X | X | X | X | X | X | X | X | X | X |
| Review data management plan and digital products plan | X | | | X | | | X | | | X | | |
| Annual reporting | | | | | | | | | | | X | |
| | | | | | | | | | | | | |
| **Research Area 2**: Chronological and Disciplinary Distributions | | | | | | | | | | | | |
| Task 3: Analysis in Two Dimensions (lead by Wu) | X | X | X | X | X | X | | | | | | |
| Task 4: How Fast OADS-URLs Disappear (Lead by Alam) | | | X | X | X | X | X | X | | | | |
| | | | | | | | | | | | | |
| **Research Area 3**: Predictive Model for Endangered OADS | | | | | | | | | | | | |
| Task 5: Building a Predictive Model for Endangered OADS (lead by Wu) | | | | | X | X | X | X | X | X | X | X |
| | | | | | | | | | | | | |
| *All investigators will collaborate and make contributions. | | | | | | | | | | | | |

| **Project Year 3** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Activities and Milestones** | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul |
| **Administrative work** | | | | | | | | | | | | |
| Advisory Board Meetings | X | | | | | | X | | | | | |
| Hire graduate research assistants | X | X | | | | | | | | | | |
| Create and maintain a project website | X | X | X | X | X | X | X | X | X | X | X | X |
| Evaluate project progress | X | X | X | X | X | X | X | X | X | X | X | X |
| Review data management plan and digital products plan | X | | | X | | | X | | | X | | |
| Annual reporting | | | | | | | | | | | X | |
| | | | | | | | | | | | | |
| **Research Area 3**: Predictive Model for Endangered OADS | | | | | | | | | | | | |
| Task 5: Building a Predictive Model for Endangered OADS (lead by Wu) | X | X | X | X | | | | | | | | |
| | | | | | | | | | | | | |
| **Research Area 4**: Integration into Digital Libraries | | | | | | | | | | | | |
| Task 6: CiteSeerX as a Testbed to Disseminate OADS-URLs (lead by Wu) | X | X | X | X | X | X | X | X | X | X | X | X |
| | | | | | | | | | | | | |
| *All investigators will collaborate and make contributions. | | | | | | | | | | | | |

# Digital Products Plan

## Type

In this project, we will develop research and non-research digital products.

The **research** digital products mainly include the following

- **Labeled data**: human labeled open-access datasets and software URLs (OADs-URLs), context sentences and metadata
- **Full data**: automatically extracted OADS-URLs, context sentences, and metadata
- **Analytical and prediction results**: analytical results in form of spreadsheets, tables, and figures
- **Log files**: search engine log files generated by CiteSeerX
- **Web archival data**: generated by web crawlers of Wayback Machine
- **Software**: machine learning models and implementations and computer programs used for preprocessing raw data (such as extracting text from PDF and configurations)

The **non-research** digital products mainly include publications (journal articles, conference proceedings, theses and dissertations, project reports), presentations (slides and posters), blog posts, and video recordings of advisory board meetings (with permissions from advisory board members).

## Availability

For **research** digital products, for reproducibility and replicability purposes, we will make most **research data** publicly available on Harvard Dataverse (https://dataverse.harvard.edu/), an open online repository for sharing, preserving, citing, exploring, and analyzing research data. Harvard Dataverse provides free unlimited storage for research data. The Web archival data will be used for experiments to study the feasibility of preserving OADS and will be available on the Internet Archive. We will make the **software** publicly available on GitHub at PI Wu's lab space (https://github.com/lamps-lab). For CiteSeerXlog files, we will anonymize them by removing the IP address and making it publicly available on Harvard Dataverse.

For **non-research** digital products, we will submit preprints on arXiv.org, a preprint service for academic papers, before the camera-ready version of papers appears on the publishers' (such as ACM and IEEE) websites. We will host a copy of publications and presentations at the Old Dominion University Digital Commons (https://digitalcommons.odu.edu/), which is an institutional repository bringing together all of a University's research under one umbrella, with an aim to preserve, provide access to, and showcase research (a letter is attached). The blog posts will be hosted (https://ws-dl.blogspot.com/) under the Web Science Digital Library (WSDL) Research Group account. The video recordings of advisory board meetings will be available at https://odumedia.mediaspace.kaltura.com and only be accessible to advisory board members, staff, and students who are involved in this project.

## Access

The **research** data will be publicly accessible on Harvard Dataverse and GitHub. The metadata on both platforms is open and findable via search engines like Google. The *research* data will be released with a Creative Commons license, specifically CC BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0), which allows users to share (copy and redistribute the material in any medium or format) and adapt (remix, transform, and build upon the material), as long as the user gives appropriate credits to the original creators without commercial purposes. The *research* software will be released with the AGPLv3 (https://www.gnu.org/licenses/agpl-3.0.html), with the goal of providing reliable and long-lived software products through collaborative, open-source software development.

The **non-research** digital products hosted by the ODU Digital Commons will be attached with an appropriate license by ODU Libraries. For example, most publications will be attached with a CC BY 4.0 license.

## Sustainability

For **research** digital products, sustainability is fulfilled by cloud-based data repository services. Specifically, the Harvard Dataverse uses Amazon Web Services and S3 and maintains a full backup on Amazon Glacier. The Harvard Faculty of Arts and Sciences maintains a separate full backup. This practice of storing data at multiple locations makes sure that the data will be highly available. The Dataverse project is financially supported by Harvard with additional support from the Alfred P. Sloan Foundation, NSF, NIH, and multiple other funding sources. GitHub is supported by its parent company Microsoft. So there is no foreseeable plan for these repositories to retire. In addition, separate digital copies will be created by the Internet Archive (IA; https://archive.org/) and made available on the IA website. Harvard Dataverse will automatically generate a standard data citation with a Digital Object Identifier (DOI).

If Dataverse or GitHub no longer exists, the team will transfer the repositories to another available platform or archive the final version in the ODU Digital Commons.

The **non-research** digital products, after ingested, become a permanent part of the ODU Digital Commons, sponsored by the University Libraries. The policy is to keep all documents as long as the Library service exists. Even if a faculty or staff member leaves the university, their work will not be withdrawn from ODU Digital Commons. The blog posts are currently owned by Google. If blogger.com no longer exists, we will transfer the blog posts to other blogger services before they are removed.

# Data Management Plan

## Overview

All Data and Software adhere to FAIR principles (DOIs, licensing, data format). All research data and software involved in this project, including source data, intermediate results, manually labeled data, automatically generated data, and software packages are part of the data management plan. The source data, including the PDFs, XMLs, and JSON files, all have open access and have been downloaded from data repositories, by following the instructions of bulk download on the data provider's protocol. The Electronic Theses and Dissertations (ETDs) have been collected by a focused web crawler. These ETDs are eligible for research purposes, covered as a case of "fair use". The labeled dataset will be created within the first six months. The full dataset will be built within the first year. The total size of OADS Repo, not including the PDFs, is expected to be less than 10GB. Data will be publicly available through Harvard Dataverse, an open online repository for sharing, preserving, citing, exploring, and analyzing research data. The Harvard Dataverse service offers unlimited storage for data repositories. To increase the sustainability of the data file, we will keep a synchronized copy on the Internet Archive, which uses their home-grown large-scale storage system called Petabox with redundant replication in more than one data centers for resilience. A permanent DOI will be generated by Harvard Dataverse. The data will be distributed under a Creative Commons License version 4.0 (CC BY-NC).

The human-labeled data and the trained machine-learning models will be hosted on the GitHub repository of the LAMP-SYS Lab, directed by PI Wu under https://github.com/lamps-lab. Software developed by us will be publicly available under the AGPLv3 license.

All papers and ETDs collected will be stored in a centralized repository server at Old Dominion University. The metadata will be stored in a relational database. Data will be processed on the high-performance computing (HPC) cluster at ODU. ODU has two HPC clusters, namely Turing and Wahab. The Turing cluster has 5600 GPU cores, 36 GPUs (a combination of Nvidia K40, K80, P100 and V100 GPUs), 34TB memory, and 180TB scratch space. The Wahab cluster has 6320 CPU cores, 72 Nvidia V100 GPUs, 60TB memory in total and 350TB storage. Both clusters are connected internally and externally using Infiniband, providing high-speed message passing between compute nodes. PI Wu has been using this cluster for research computation in many projects. At Internet Archive, the Wayback Machine APIs will be used to preserve web resources that are found live, but are not archived anywhere, hence considered endangered. Web archive data is stored in WARC files at the Internet Archive and is made accessible via the Wayback Machine.

## Types of Data

The data types and instances involved in this project are shown in Table 1.

## Data Standards

To make the research data machine-readable, we will adopt standard open data formats such as CSV, XML, JSON, and SQL formats. CSV will be used for tabular data. JSON and XML formats are used for structured full text, metadata, and semi-structured features extracted from PDFs. Throughout the work, all file encodings conform to UTF8. SQL files will be used for MySQL, a relational database management system. Because of the use of open and widely adopted standards, data migration will not be difficult should the need for media, format, or hardware changes arise. The Linux-based platform for the storage server allows easy migration to other systems.

## Data Access Policies and Redistribution

All source data obey the access policies of the original data distributors. Manually labeled ground truth data and databases will be publicly available via GitHub. The full dataset will be shared via Harvard Dataverse. Any user identifiers will be removed before the data is distributed. Data that cannot be shared, such as the full text

**Table 1:** Data types and instances involved in this project.

| Data Type | Description | Sources | Storage |
|---|---|---|---|
| Raw | PDFs | Focused web crawling | Local file system |
| Metadata | Publication year, disciplines etc. | API, OAI-PMH | Local database |
| Intermediate | Text extracted from PDFs | PDF to text converters and OCR | Local file system |
| Labeled data | Labeled URLs, context sentences, and properties at various levels | Human labeling | Local + GitHub |
| Full dataset | Automatically extracted URLs, context sentences, etc. | Information extraction with machine learning models | Local + Harvard Dataverse + Internet Archive |
| CiteSeerX linking data | Classified OADS-URLs linking to CiteSeerX papers | Automatic matching | CiteSeerX database |
| Software | Machine learning models and training and testing data | Software engineering | Local + GitHub |
| Web Archive | WARC files and its derivatives resulting from web archiving | Web crawlers | Wayback Machine |

of ETDs, will remain on a limited-access and password-protected server. The OADS-URLs that match CiteSeerX papers will be available on CiteSeerX's website.

### Data Storage Durability and Data Integrity

All the research data used for data processing will be stored on the ODU HPC server. Linux operating systems (OS) will be installed on these servers for compatibility, security, and efficiency. The ODU HPC clusters are highly redundant with clustered head nodes and a dedicated login node.

Data integrity is guaranteed through redundancy on disk, file system, and software levels. We will keep a single copy of data that can be regained (such as PDFs) and at least two copies of data that are not publicly accessible or require considerable computational effort (such as labeled and the full dataset). To avoid accidental data loss, we keep one copy of key data locally and in the cloud. Firewalls are setup on the servers so access to crucial production servers is restricted to a handful of servers and workstations within the local network. Mission-critical servers are only accessible by people through two-factor authentication.

### Documentation

All published data will be attached to a datasheet describing comprehensive issues related to the motivation, composition, collection process, recommended uses, maintenance, etc. We will follow the template provided by Gebru et al. (2018) titled "Datasheet for Datasets". This template has been widely used in AI/ML and related communities. PI Wu had experience of editing such a document with his recent "DeepPatent2" dataset.

### Collaborations

The collaborators from the Internet Archive and Virginia Tech will strictly follow the data management plan set by the leading institute. In addition, the teams will share all source codes via the central GitHub repository provided by the ODU team. The teams will share research data with each other via public cloud, such as Google Cloud, or via `ssh` between remote servers. The three institutions: ODU, Internet Archive, and Virginia Tech use VPNs to protect data security and control access to servers. The data management plan will be reviewed collaboratively on a quarterly basis.