Toward Robust URL Extraction for Open Science: A Study of arXiv File Formats and Temporal Trends

Presented By:

Rochana R. Obadage[1]

rochana@cs.odu.edu

Co-Authors:

Lamia Salsabil[1], Sawood Alam[2], William A. Ingram[3], Bipasha Banarjee[3], Edward A. Fox[3] and Jian Wu[1]

[1]Old Dominion University, Norfolk VA, USA
[2]Internet Archive, San Francisco, CA, USA
[3]Virginia Tech, Blacksburg, VA, USA

Web Archiving and Digital Libraries Workshop, September 15, 2025

Paper ID : **2113**

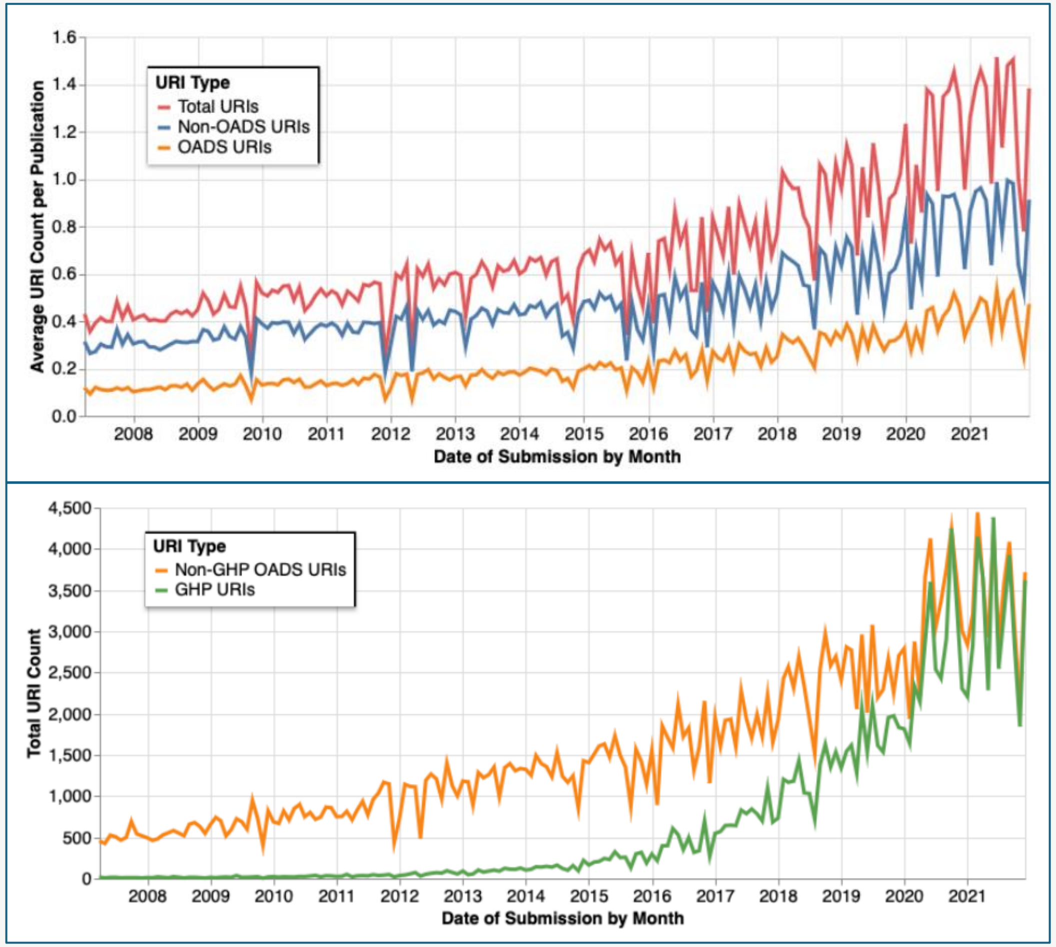
Use of URLs in Scholarly Communication

- URLs are essential connectors in scholarly publishing, linking papers to datasets, software, supplementary materials, and external resources

- They support reproducibility by enabling researchers to access the exact materials used in prior studies

- As open science grows, URLs also play a vital role in meta-analysis, large-scale data mining, and web archiving

- However, the persistence of these links is fragile, threatening the long-term accessibility of scientific knowledge



Increasing use of URIs in scholarly articles: Figures adapted from Escamilla et al., [10]

Motivation & Problem Statement

- Reproducibility depends on persistent access to datasets and software referenced in scholarly papers
- URLs in papers provide essential links, but link rot and extraction challenges cause many URLs to disappear or be missed
- Incomplete URL extraction limits reproducibility and digital preservation, especially for long-term archiving
- **Our ongoing initiative:** Preserving Open Access Datasets and Software – building the largest corpus of open-access dataset and software URLs from scholarly literature
    - Currently leveraging 2.3M+ full-text arXiv papers, S2ORC 6.3M+ (Text), PubMed 6.9M+ (XML, Text)
- This preliminary study is part of that larger effort, focusing on how different input formats influence URL extraction from scholarly articles

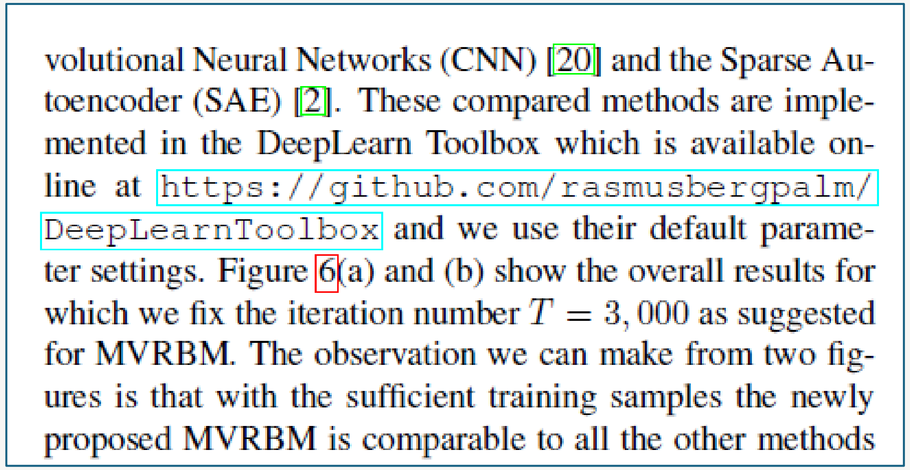
Background

arXiv

- A major **open-access digital library** with 2.3M+ full-text papers since 1991, spanning multiple disciplines
- Acts as both an **early dissemination platform** and a **long-term digital archive**
- File Formats
  - Directly provided: PDF, LaTeX source, PostScript, HTML (partially)
  - Indirectly derived : XML (via GROBID), HTML (via LaTeXML)

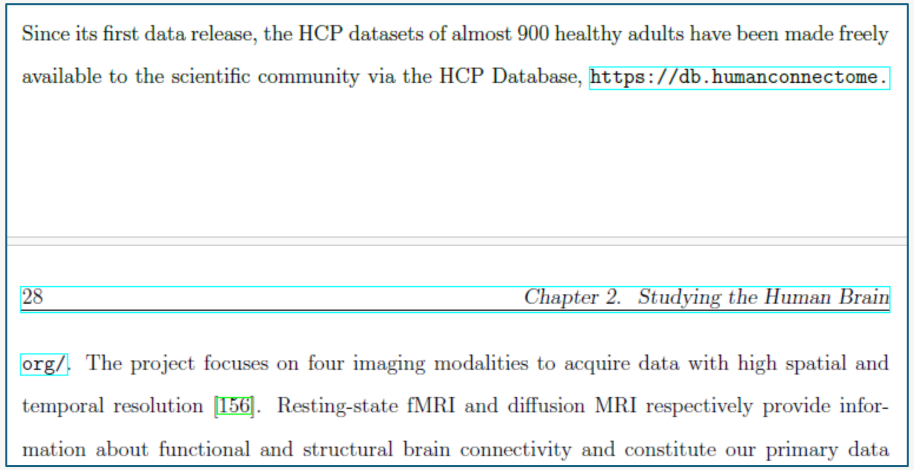
Extraction challenges

- PDFs break text flow and split URLs.
- Annotation-layer URLs often ignored.
- Reliable URL extraction across formats is still unsolved.



volutional Neural Networks (CNN) [20] and the Sparse Autoencoder (SAE) [2]. These compared methods are implemented in the DeepLearn Toolbox which is available online at https://github.com/rasmusbergpalm/DeepLearnToolbox and we use their default parameter settings. Figure 6(a) and (b) show the overall results for which we fix the iteration number $T = 3,000$ as suggested for MVRBM. The observation we can make from two figures is that with the sufficient training samples the newly proposed MVRBM is comparable to all the other methods

A URL split into two lines in PDF (https://arxiv.org/pdf/1802.06772)



Since its first data release, the HCP datasets of almost 900 healthy adults have been made freely available to the scientific community via the HCP Database, https://db.humanconnectome.

28 *Chapter 2. Studying the Human Brain*

org/. The project focuses on four imaging modalities to acquire data with high spatial and temporal resolution [156]. Resting-state fMRI and diffusion MRI respectively provide information about functional and structural brain connectivity and constitute our primary data

A URL split into two pages in PDF (https://arxiv.org/pdf/1702.05374)

Related Work

Existing Approaches to URL Extraction

- Most prior work focuses on a **single file format**, primarily PDFs or preprocessed Text.
- Large-scale projects (e.g., **S2ORC**, **Semantic Scholar**) rely on regex and heuristic filters applied to PDF/Text.
- Tools such as **GROBID** and **PyMuPDF** are optimized for PDF-based extraction, targeting structured metadata and references.
- GROBID, for example, uses PDFBox for text extraction before applying ML models to parse document structure.

Benefits of Structured Formats (HTML/XML)

- HTML and XML retain **semantic structure** of documents, making them more "machine-friendly."
- Studies show these formats **preserve URL fidelity** better than PDF-derived text.
- Prior works using HTML/XML do **not focus on URL extraction directly**, nor do they compare formats systematically.

Link Decay and Reproducibility Challenges

- Academic links **decay over time**, with rates varying by domain and hosting platform.
- Prior studies (e.g., Klein et al.,[15] Hennessey et al. [13]) show widespread **link rot** in scholarly works.
- A 2025 GitHub repository study [6] found **10% had no archived version,** and even archived pages were often incomplete.
- Hybrid approaches (e.g., Escamilla et al. [10]) can detect open-access URIs across diverse platforms but are **domain-specific** or use **off-the-shelf tools.**

Gaps in Current Literature

- **Over-reliance on PDF text extraction:** Broken text flow causes missed or truncated URLs.
- **Lack of systematic benchmarks** for evaluating URL extraction across multiple input formats.
- Without benchmarks, it is difficult to **assess system accuracy** or **develop improvements** for scholarly URL extraction.

Research Objectives / Contributions

Research Objectives:

- Find out which input format, format combination works best for URL extractions from Scholarly Papers
- Evaluate URL extraction performance across multiple arXiv file formats
- Build a pilot dataset with ground-truth URLs
- Analyze temporal trends in arXiv URL usage (1992–2024)

Contributions:

- A comparative evaluation of URL extraction performance across four file formats (Text, LaTeX, HTML, and XML), identifying which formats yield the best performance
- A longitudinal analysis of URLs in arXiv papers, revealing how the overall usage of URLs in a yearly sample has evolved over three decades.

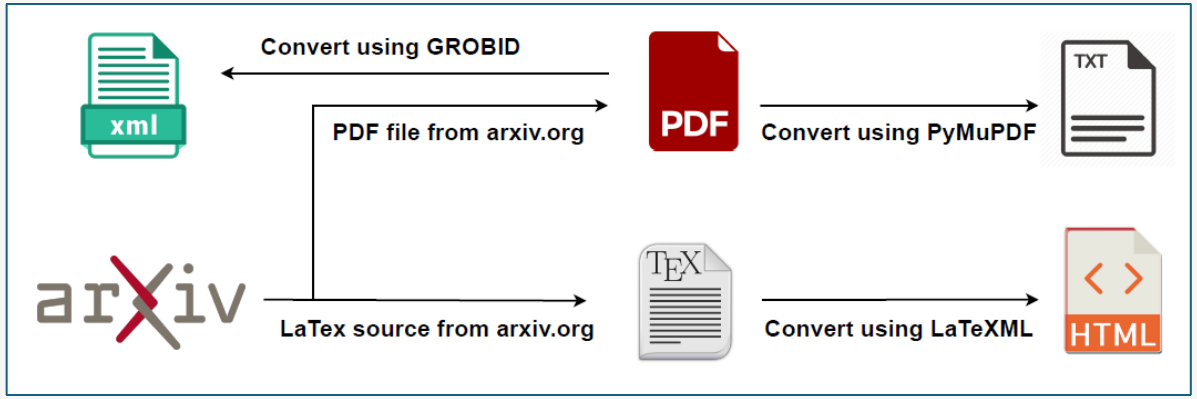
Dataset Preparation

- **Sample:** 1,161 papers (1991–2024), ~3 from each "year-month" stratum

- **Coverage:** Only 60 papers had URLs extracted from all 3 sources (Text, LaTeX, HTML)

- **Pilot dataset:** Randomly selected 10 papers with all 5 formats

- **Ground truth:** 87 manually verified URLs from PDFs



File format coverage and conversion tools used for the randomly selected papers

Format	Conversion Tool	# Papers	# Papers with Extracted URLs
PDF	–	1,161	–
Text	PyMuPDF	1,161	260
LaTeX	–	726	252
HTML	LaTeXML	204	134
XML	GROBID	60	60

Methodology

Ground Truth - Extended (Superset)

- Manual PDF inspection (87 URLs) + URL-like candidates

URL Extraction Approach

- **Text**: regular expression (regex) patterns
- **LaTeX**: regex + **\*url***, **\*urladdr*** anchors in *.tex* | *.bbl* files
- **HTML**: using anchor ***<a>*** tags and ***href*** attribute
- **XML** (TEI): extract elements with ***target*** attribute

```
<ref target="https://github.com/kermitt2/grobid"/>
```

Tools

- PyMuPDF (v1.24.13) - https://pypi.org/project/PyMuPDF/
- LaTeXML (v0.8.8) - https://github.com/brucemiller/LaTeXML
- GROBID (v0.8.1) - https://github.com/kermitt2/grobid

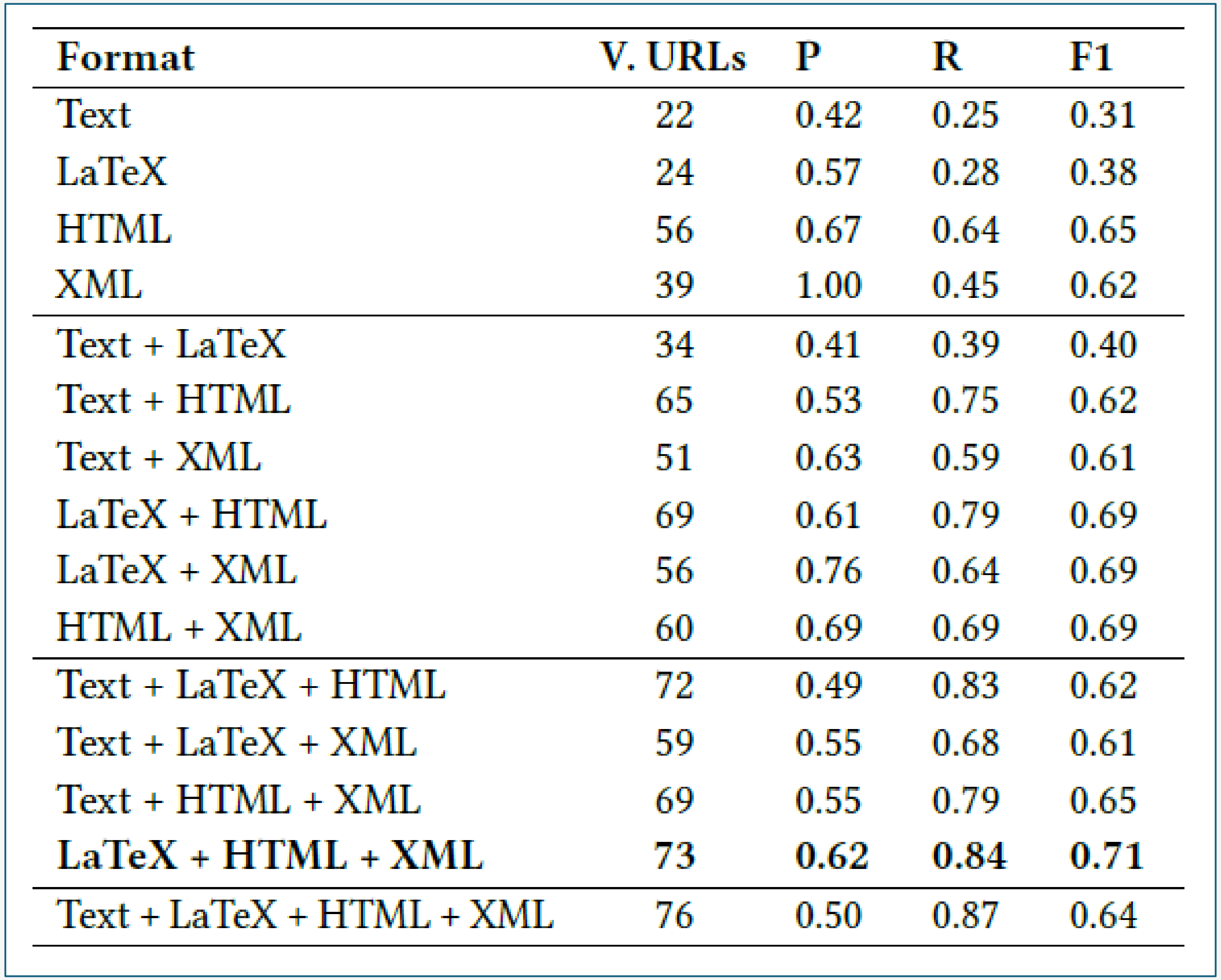
URL-like candidates (URL strings)

- Valid URLs (appears in the PDF version)
- Invalid URLs (not found in the PDF)
- Broken URLs (partially extracted due to line or page breaks)

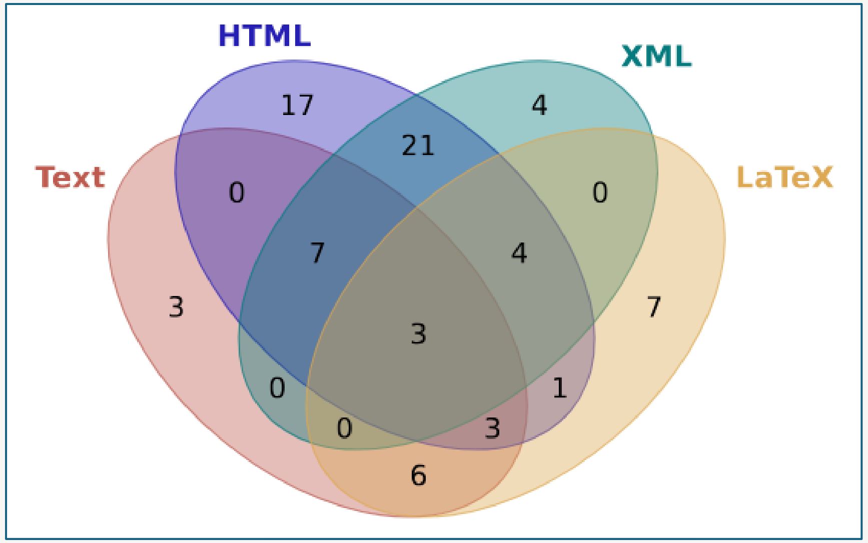
Evaluation metrics

- $\text{Precision} = \dfrac{\#\ \text{extracted valid URLs}}{\#\ \text{total extracted URL strings}}$
- $\text{Recall} = \dfrac{\#\ \text{extracted valid URLs}}{\#\ \text{total valid URLs}}$

Results – Single Formats & Format Combinations



Format	V. URLs	P	R	F1
Text	22	0.42	0.25	0.31
LaTeX	24	0.57	0.28	0.38
HTML	56	0.67	0.64	0.65
XML	39	1.00	0.45	0.62
Text + LaTeX	34	0.41	0.39	0.40
Text + HTML	65	0.53	0.75	0.62
Text + XML	51	0.63	0.59	0.61
LaTeX + HTML	69	0.61	0.79	0.69
LaTeX + XML	56	0.76	0.64	0.69
HTML + XML	60	0.69	0.69	0.69
Text + LaTeX + HTML	72	0.49	0.83	0.62
Text + LaTeX + XML	59	0.55	0.68	0.61
Text + HTML + XML	69	0.55	0.79	0.65
LaTeX + HTML + XML	**73**	**0.62**	**0.84**	**0.71**
Text + LaTeX + HTML + XML	76	0.50	0.87	0.64



Overlap of extracted valid URLs across Text, LaTeX, HTML, andXML

Evaluation metrics

- **Precision** $= \dfrac{\text{\# extracted valid URLs}}{\text{\# total extracted URL strings}}$

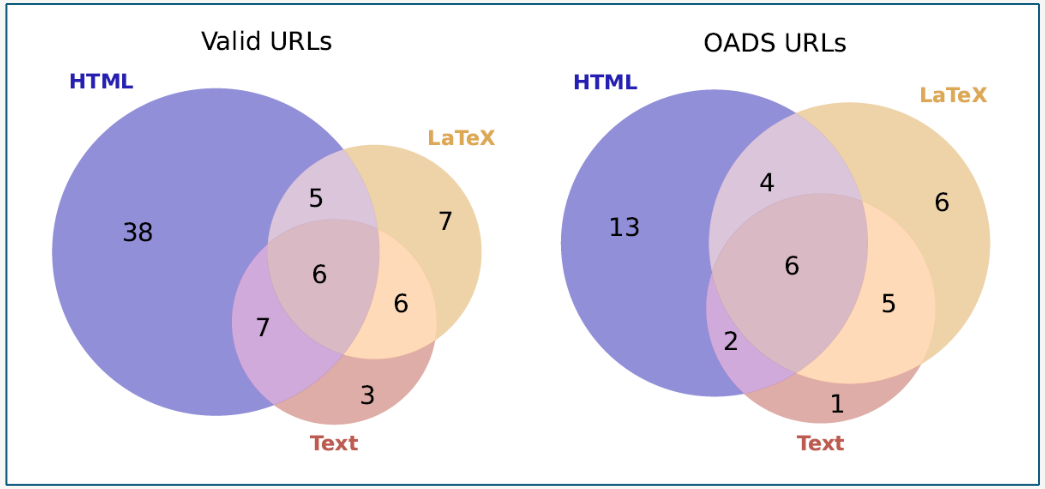
- **Recall** $= \dfrac{\text{\# extracted valid URLs}}{\text{\# total valid URLs}}$

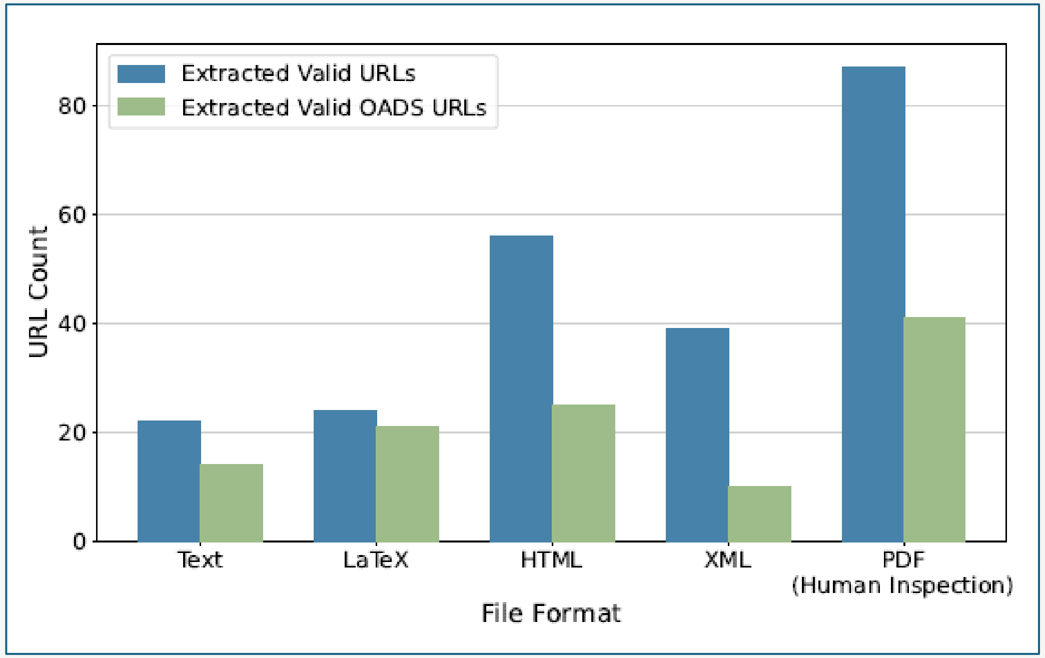
OADS URLs

Open Access Datasets & Software (**OADS**) URLs: critical for reproducibility

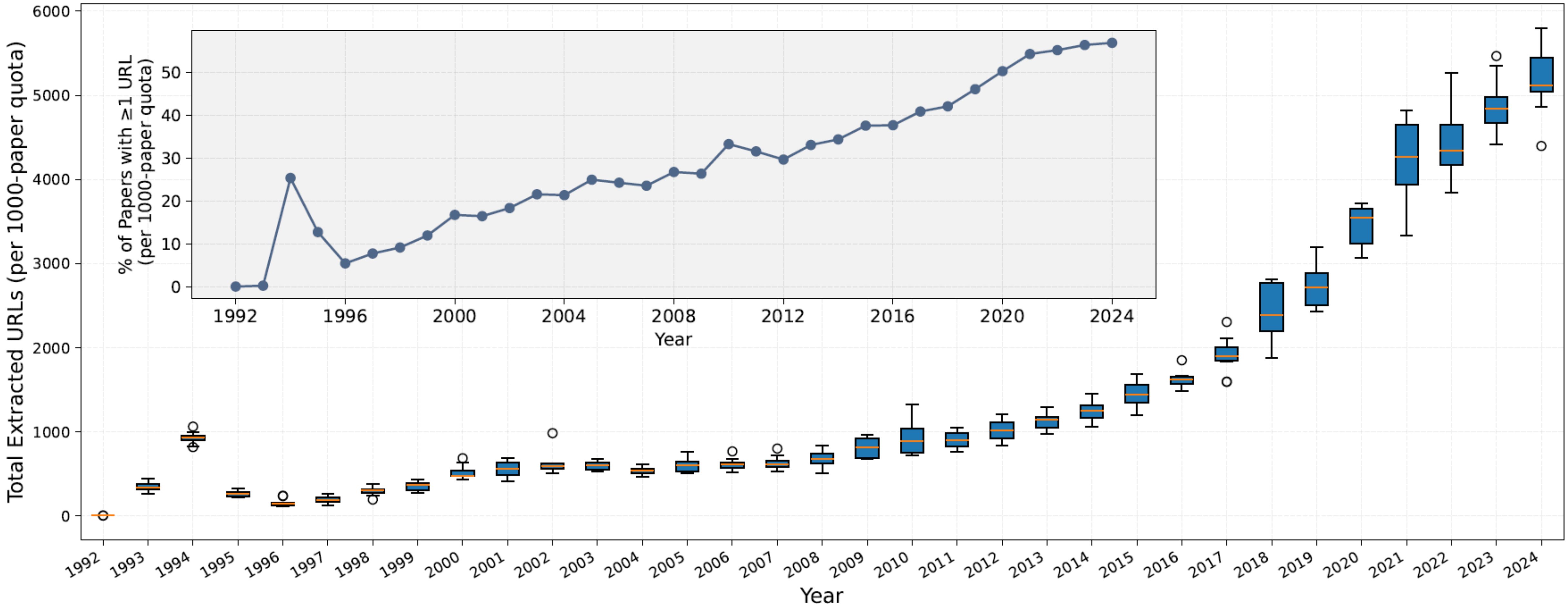
Extraction results

- **41**/87 ground-truth URLs were OADS
- **HTML:** most OADS URLs (25/41)
- **LaTeX:** highest proportion (87.5%) of valid URLs relative to the total extracted per format
- Overlap analysis: each format misses some; combining formats increases completeness





Temporal Trends



Composite distribution of extracted URLs from arXiv papers. The boxplot for each year is obtained by 10 random draws of 1,000 papers with replacement from all papers published in that year (461,418 URLs from ~330,00 papers). Inset: Percentage of papers containing at least one URL, based on a single random sample of 1,000 papers per year.

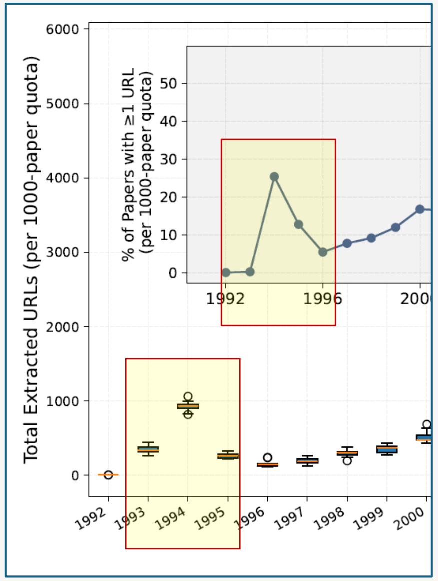
Discussion

Key findings

- No single format is sufficient — input format strongly affects extraction quality.
- Structured formats (HTML, LaTeX, XML) preserve fidelity better than Text/PDF.
- Multi-format extraction **improves coverage by ~39–51%,** though at the cost of more false positives.
- **Temporal trends** show exponential growth in URL usage (0.02% in 1990s → ~55% today).
- **1994 anomaly:** surge due to early web adoption in Computer Science & Physics papers.

Challenges

- Combining multiple formats increases false positives.
- **Filtering is needed** to reduce noise and account for **broken or invalid links.**



Limitations

- **Small pilot dataset:** only 10 papers with all formats
- Not all arXiv papers consistently provide LaTeX/HTML/XML
- Temporal analysis based on Text-only extraction
- Heuristic regex-based methods only; ML-based methods not yet tested

Conclusions & Future Work

Conclusions

- **HTML, XML,** and **LaTeX** outperform Text for URL extraction
- Combining formats ensures more robust, complete coverage
- Multi-format extraction is vital for capturing OADS URLs
- Temporal analysis shows rapid growth of URL usage over 30 years

Future Work

- Expand human annotation to more papers; build robust benchmark dataset
- Extend study to other repositories (S2ORC, PubMed)
- Study link longevity and preservation coverage
- Build a large corpus of OADS URLs to support preservation efforts

Summary

1. Multi-format input file URL extraction significantly outperforms single formats, improving URL coverage by ~39% – 51% and demonstrating that no individual format is sufficient on its own

2. Structured formats such as **HTML**, **XML**, and **LaTeX** better preserve URL fidelity, while plain text often leads to truncation or broken links due to disrupted text flow

3. Longitudinal analysis (1992–2024) shows a rapid growth of URL usage in scholarly communication, highlighting how the role of URLs in scholarly communication has evolved over three decades

Resources / Acknowledgments

Dataset & Notebooks

https://github.com/lamps-lab/arxiv-urls

Funding support

Partially funded by the Institute of Museum and Library Services through the Grant for **Preserving Endangered Open Access Datasets and Software** project.

References

[1] 2008–2025. GROBID. https://github.com/kermitt2/grobid. swh:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c

[2] Kehinde Ajayi, Muntabir Hasan Choudhury, Sarah M. Rajtmajer, and Jian Wu. 2023. A Study on Reproducibility and Replicability of Table Structure Recognition Methods. https://doi.org/10.1007/978-3-031-41679-8_1, 3–19 pages.

[3] arXiv. 2025. arXiv.org: e-Print archive for Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, and Statistics. https://arxiv.org . Accessed: 2025-07-01.

[4] Deyan Ginev Bruce R. Miller. 2004. LaTeXML A LaTeX to XML/HTML/MathMLConverter. https://github.com/brucemiller/LaTeXML . [Accessed 27-06-2025].

[5] Duy Duc An Bui, Guilherme Del Fiol, and Siddhartha Jonnalagadda. 2016. PDF text classification to leverage information extraction from publication reports. Journal of Biomedical Informatics 61 (2016), 141–148. doi:10.1016/j.jbi.2016.03.026

[6] David Calano, Michael Nelson, and Michele Weigle. 2025. GitHub Repository Complexity Leads to Diminished Web Archive Availability. In Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25). Association for Computing Machinery, New York, NY, USA, 449–459. doi:10.1145/3717867.3717920

[7] CERN. 2024. A short history of the Web. https://home.cern/science/computing/birth-web/short-history-web Accessed: 2025-07-01.

[8] Sandra L De Groote, Mary Shultz, and Deborah D Blecic. 2014. Informationseeking behavior and the use of online resources: a snapshot of current health sciences faculty. J Med Libr Assoc 102, 3 (July 2014), 169–176.

[9] Jingcheng Du, Dong Wang, Bin Lin, Long He, Liang-Chin Huang, Jingqi Wang, Frank J Manion, Yeran Li, Nicole Cossrow, and Lixia Yao. 2025. Use of deep learning-based NLP models for full-text data elements extraction for systematic literature review tasks. Scientific Reports 15, 1 (June 2025), 19379.

[10] Emily Escamilla, Lamia Salsabil, Martin Klein, Jian Wu, Michele C. Weigle, and Michael L. Nelson. 2023. It's Not Just GitHub: Identifying Data and Software Sources Included in Publications. In Linking Theory and Practice of Digital Libraries, Omar Alonso, Helena Cousijn, Gianmaria Silvello, Mónica Marrero, Carla Teixeira Lopes, and Stefano Marchesin (Eds.). Springer Nature Switzerland, Cham, 195–206.

[11] Rodney Kinney et al. 2025. The Semantic Scholar Open Data Platform. arXiv:2301.10140 [cs.DL] https://arxiv.org/abs/2301.10140

[12] Paul Ginsparg. 2011. It was twenty years ago today ... arXiv:1108.2700 [cs.DL] https://arxiv.org/abs/1108.2700

References

[13] Jason Hennessey and Steven Xijin Ge. 2013. A cross disciplinary study of link decay and the effectiveness of mitigation techniques. BMC Bioinformatics 14, 14 (Oct. 2013), S5.

[14] Susan Howell and Amber Burtis. 2022. The continued problem of URL decay: an updated analysis of health care management journal citations. J Med Libr Assoc 110, 4 (Oct. 2022), 463–470.

[15] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLOS ONE 9, 12 (12 2014), 1–39. doi:10.1371/journal.pone.0115253

[16] Viktor Lakic, Luca Rossetto, and Abraham Bernstein. 2023. Link-Rot in Web-Sourced Multimedia Datasets. In MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I (Bergen, Norway). Springer-Verlag, Berlin, Heidelberg, 476–488. doi:10.1007/978-3-031-27077-2_37

[17] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 4969–4983. doi:10.18653/v1/2020.acl-main.447

[18] Zara Nasar, SyedWaqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. Scientometrics 117, 3 (Dec. 2018), 1931–1990. doi:10.1007/s11192-018-2921-5

[19] Pavel Panchekha and Chris Harrelson. 2025. History of the Web. In Web Browser Engineering. Oxford University Press. doi:10.1093/9780198913887.003.0003 arXiv:https://academic.oup.com/book/0/chapter/498097450/chapterpdf/61200141/isbn-9780198913887-book-part-2.pdf

[20] Lamia Salsabil, Jian Wu, Muntabir Hasan Choudhury, William A. Ingram, Edward A. Fox, Sarah M. Rajtmajer, and C. Lee Giles. 2022. A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software. In Companion Proceedings of the Web Conference 2022 (Virtual Event, Lyon, France) (WWW'22), 784–788. doi:10.1145/3487553.3524658

[21] Julian Smith. 2016. PyMuPDF. https://pypi.org/project/PyMuPDF/. [Accessed 27-06-2025].

[22] Ke Zhou, Richard Tobin, and Claire Grover. 2014. Extraction and analysis of referenced web links in large-scale scholarly articles. In IEEE/ACM Joint Conference on Digital Libraries. 451–452. doi:10.1109/JCDL.2014.6970220