LA-UR-22-29462

Approved for public release; distribution is unlimited.

Title:	PDF Server: "Robustifying" URL References in PDF Documents
Author(s):	Jayawardana, Yasith Cain, Brian J. Klein, Martin Jones, Shawn Morgan
Intended for:	Mini Summer Student Symposium 2022, 2022-08-25 (Los Alamos, New Mexico, United States)
Issued:	2022-09-12

PDF Server: "Robustifying" URL References in PDF Documents

Yasith Jayawardana
Proto-team, SRO-RL
Supervisor: Martin Klein

Mini Summer Student Symposium 2022

PhD Student
Department of Computer Science, Old Dominion University
Norfolk, Virginia

Los Alamos NATIONAL LABORATORY

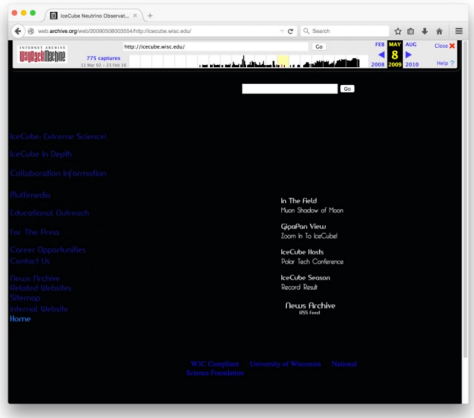
WSDL Old Dominion University

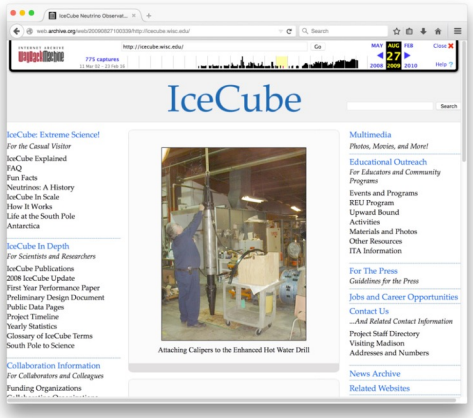
ODU

Links on the Web Break all the Time

Link Rot: URL references may **break** over time

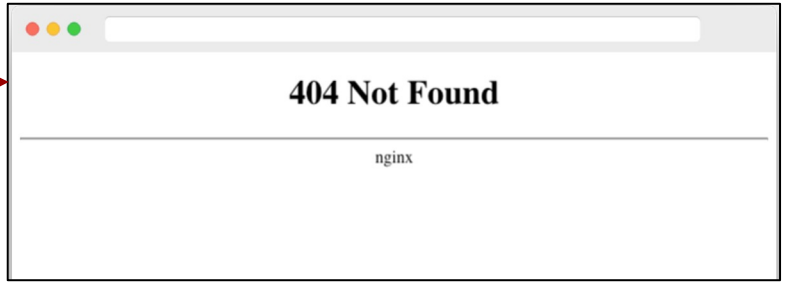
Content Drift: URL contents may **change** over time

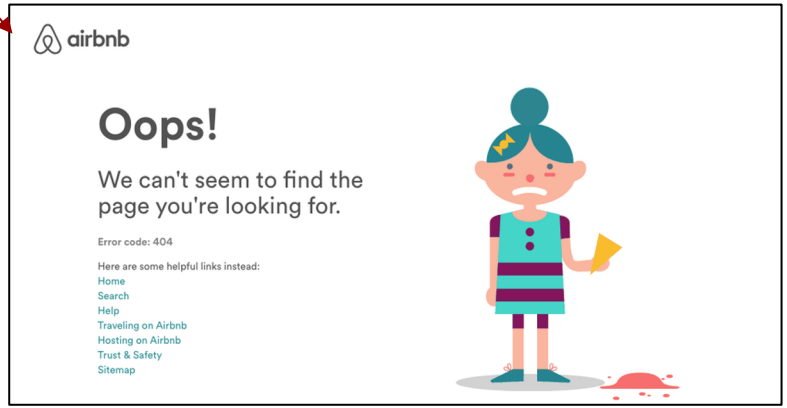




Hard 404

Soft 404





"Robustifying" URLs to Combat **Link Rot** & **Content Drift**

Given an HTML hyperlink,

```
<a href="https://creativecommons.org/licenses/by/4.0">https://creativecommons.org/licenses/by/4.0</a>
```

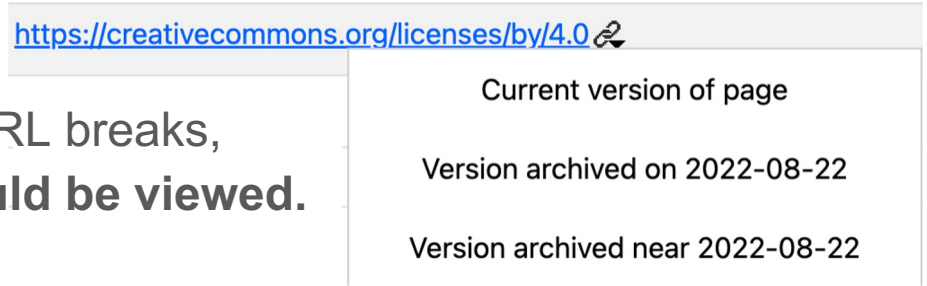
Step 1 - Create a snapshot of the referenced content in a public Web Archive

e.g., archive.is, archive.org

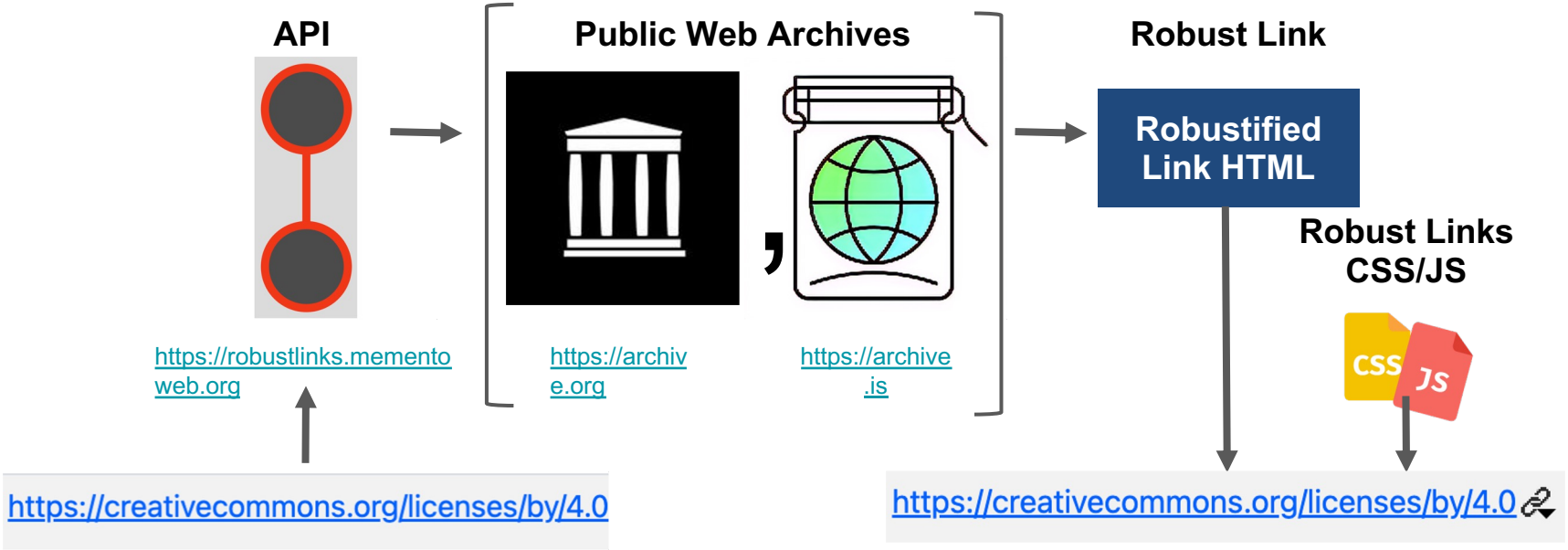
Step 2 - Embed information about this snapshot into the HTML

```
<a href="https://creativecommons.org/licenses/by/4.0" data-versionurl="https://archive.ph/wip/BUdBd"
data-versiondate="2022-08-22">https://creativecommons.org/licenses/by/4.0</a>
```

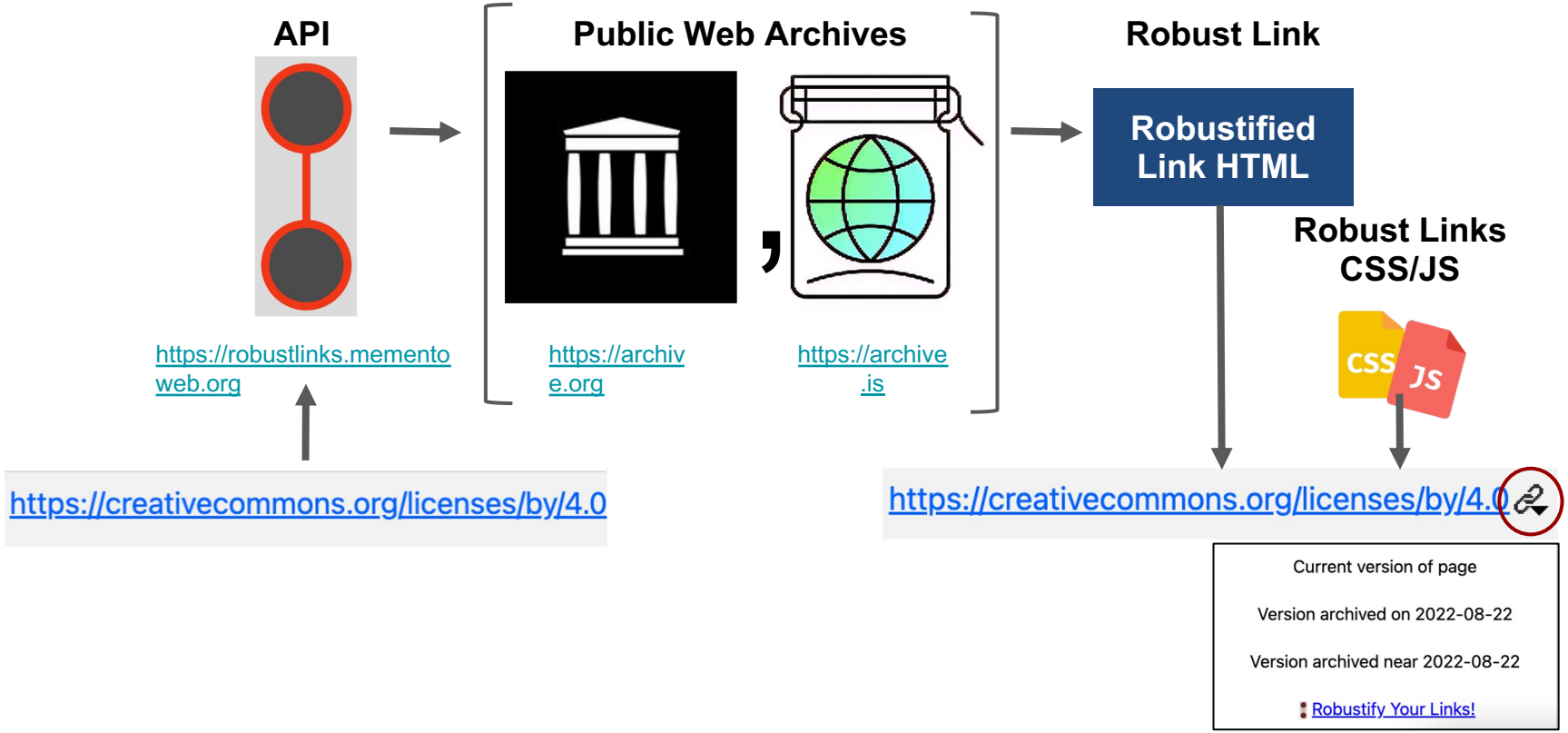
Step 3 - Show users that if/when that URL breaks, **there's still an archived copy that could be viewed.**



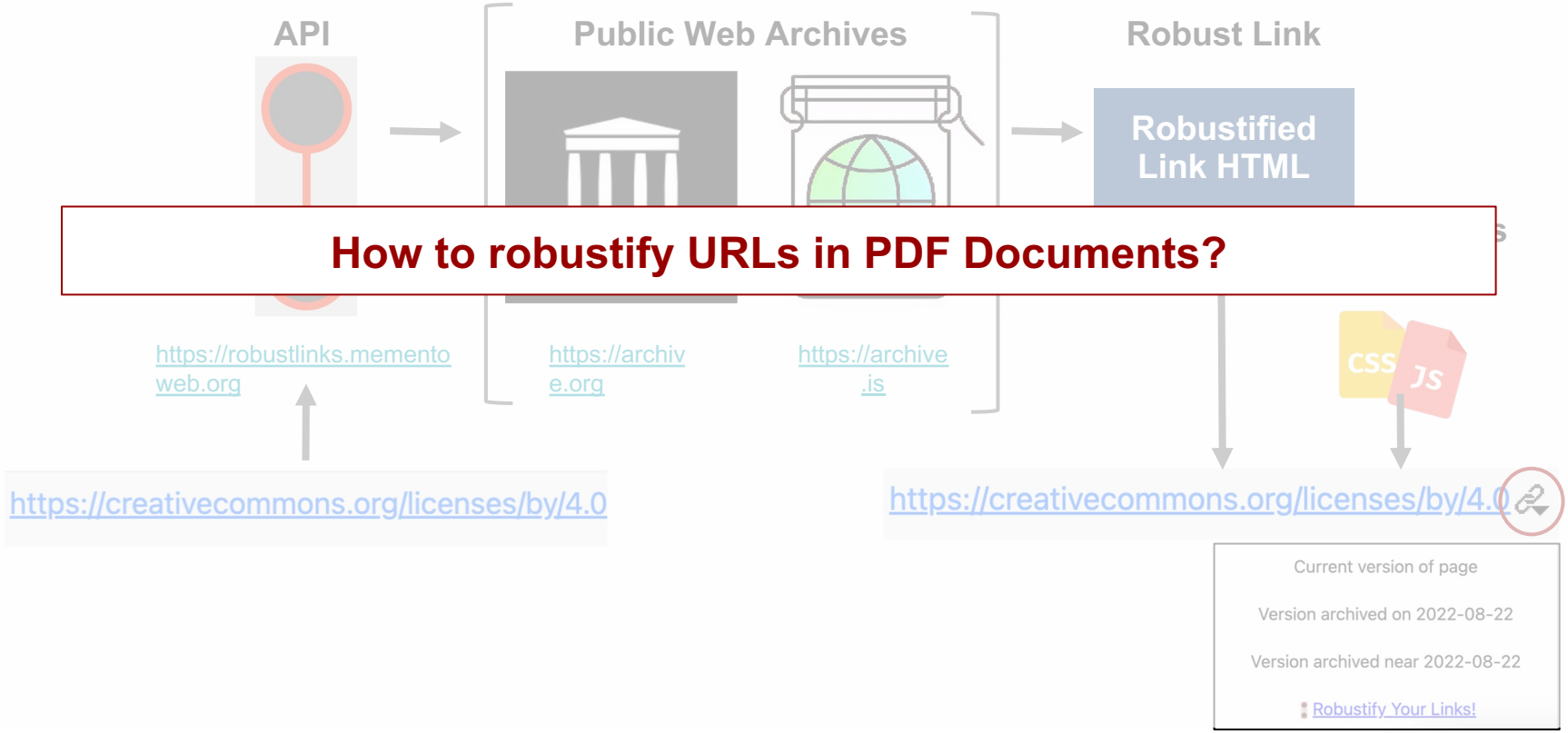
Robustifying URLs using the Robust Links Service



Robustifying URLs using the Robust Links Service



Robustifying URLs using the Robust Links Service



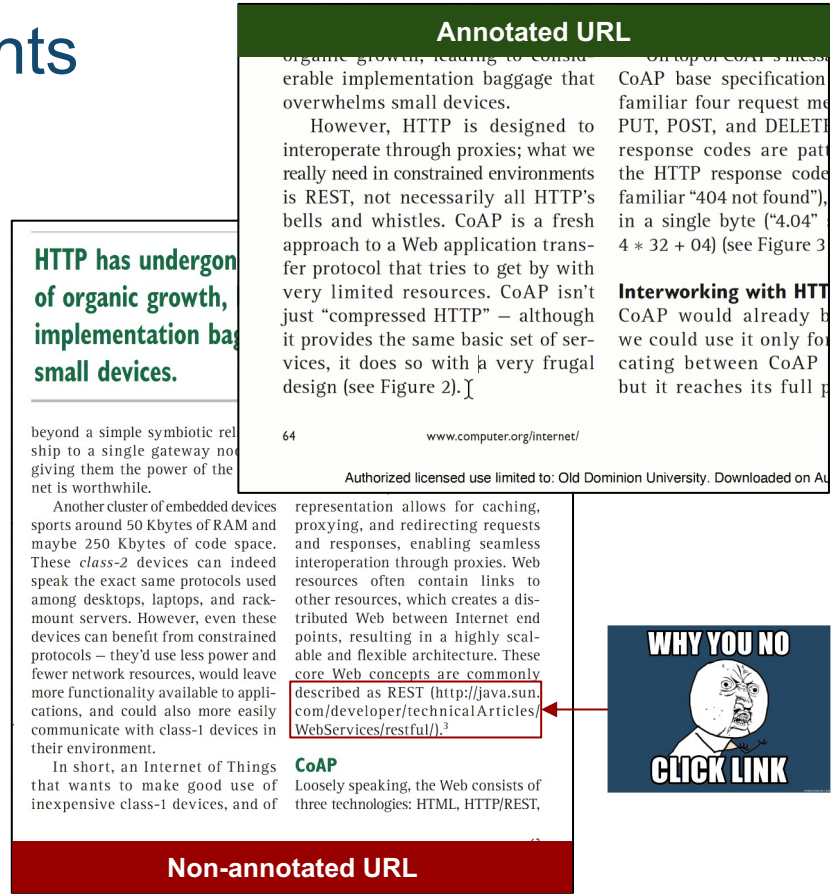
Extracting URLs in PDF Documents

(My Summer 2021 Internship Project)

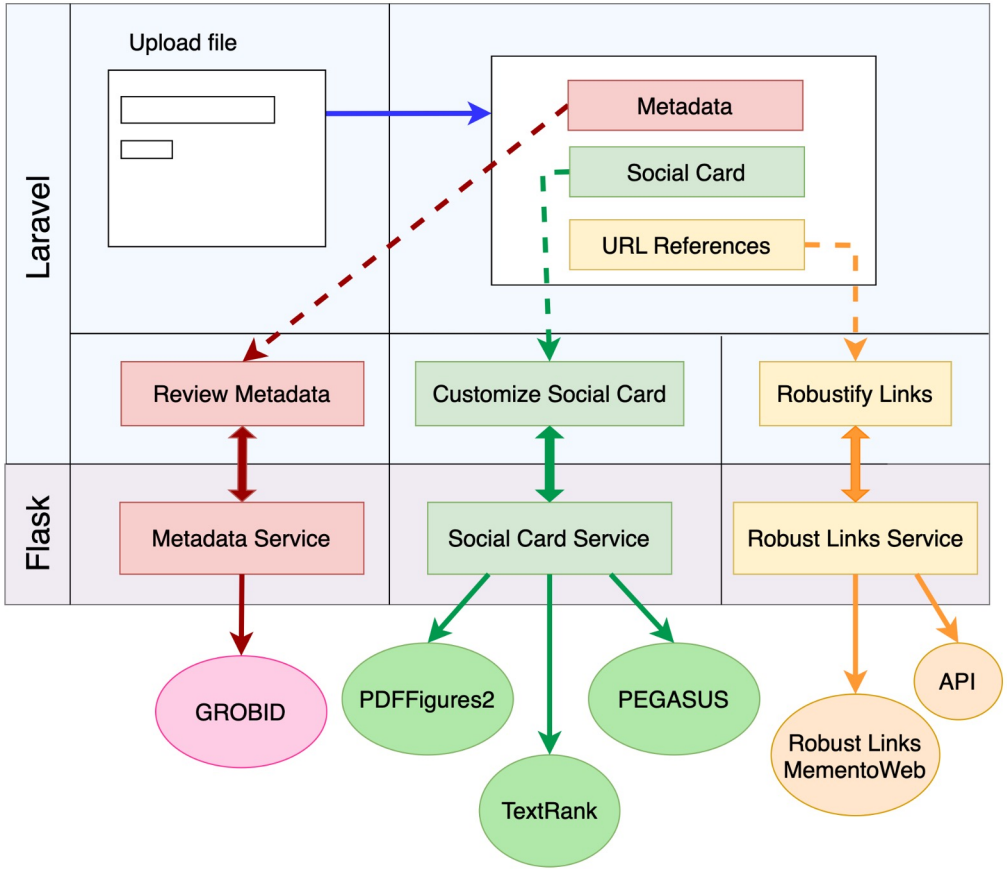
- URLs may (or may not) be annotated
  - **Annotated URLs** are interactive and **easy to extract**
  - **Non-annotated URLs** are text-like and **hard to extract**

- We frequently see non-annotated URLs.
  It's not straightforward to extract them
  - Newline separated
  - Expanded hyphens
  - URL spread across two columns

- Tools we used
  - Extracting annotated URLs - **PyPDF2**
  - Extracting non-annotated URLs - **PDFIUM**
    - PDFIUM is used in Chrome to render PDF documents
    - **It identifies URLs in PDF text remarkably well**



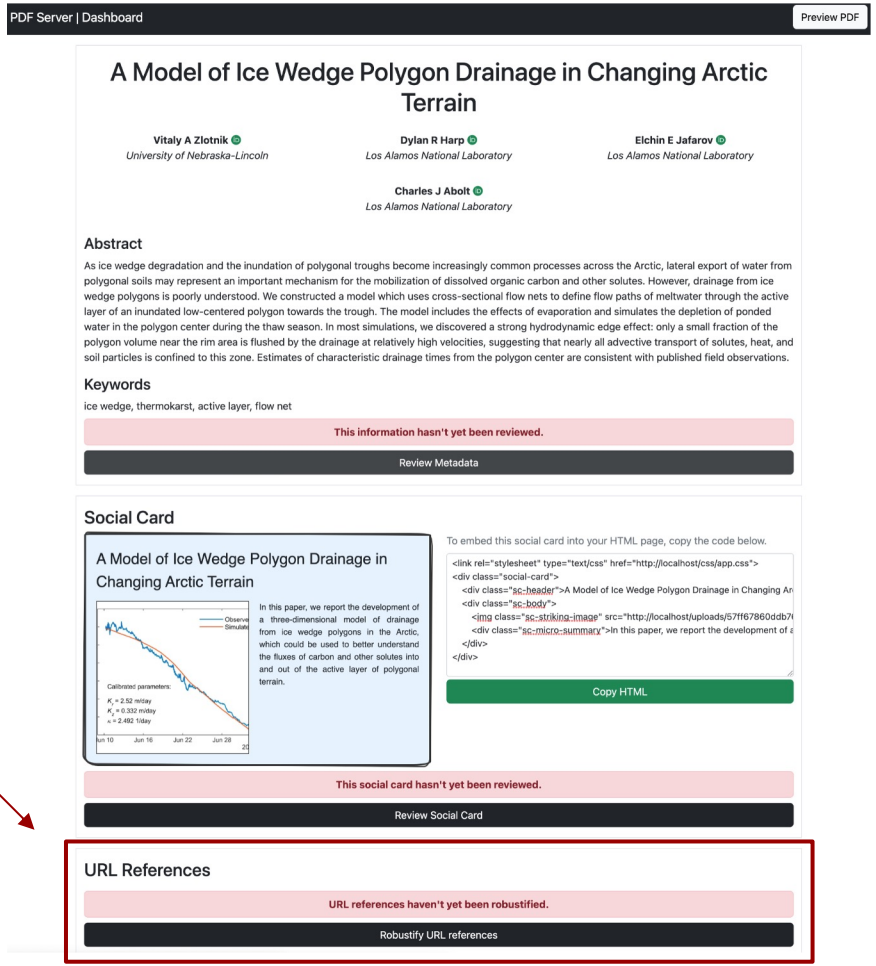
Overall Architecture of PDF Server



Robustify Links on PDF Server

<u>Idea:</u> Allow users to select useful URLs from automatically extracted URLs.

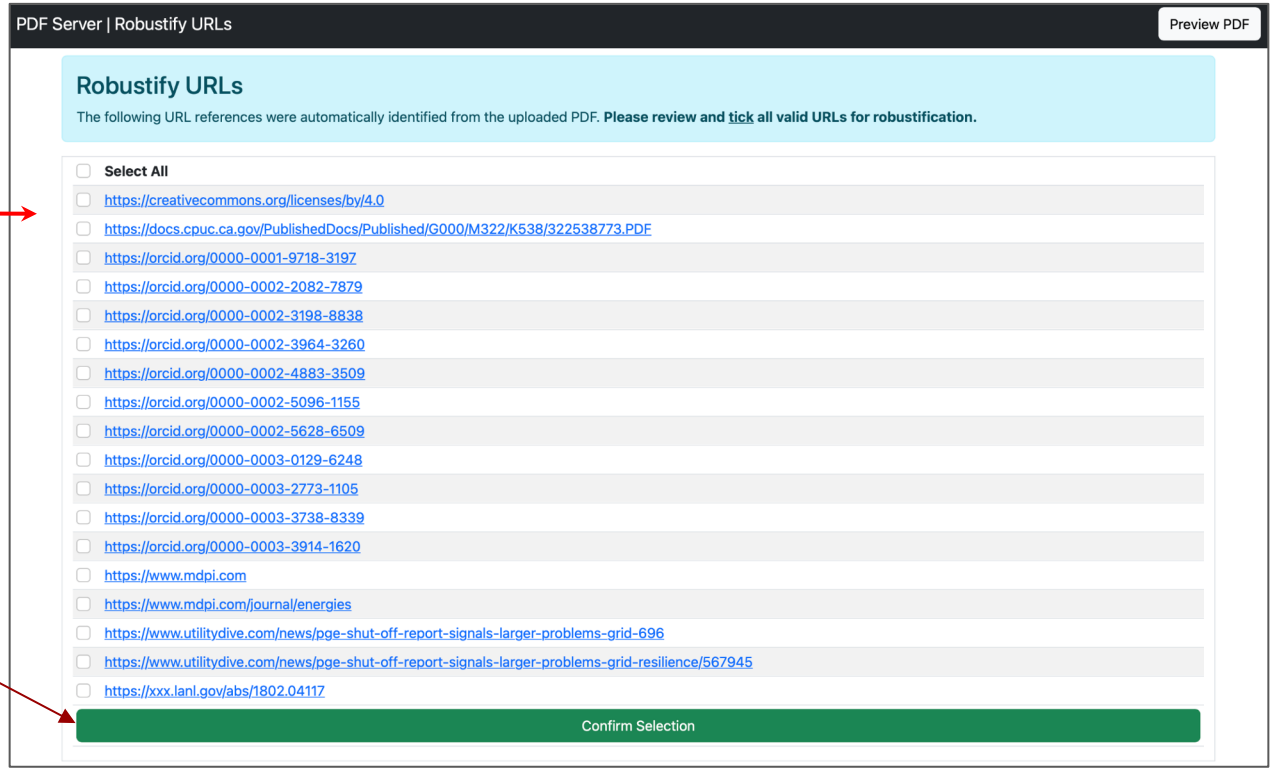
A warning message indicates whether any URLs have been selected yet.



Selecting URLs from Automatically Extracted URLs

The user can choose which URLs to robustify, or select all extracted URLs using "Select All."

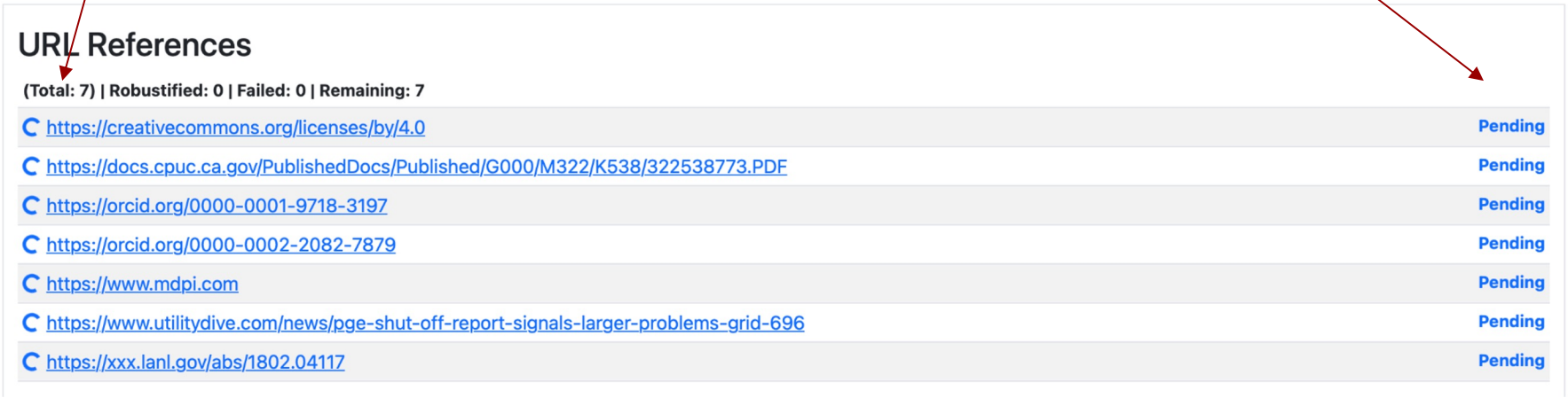
After selecting URLs, user confirms the selection



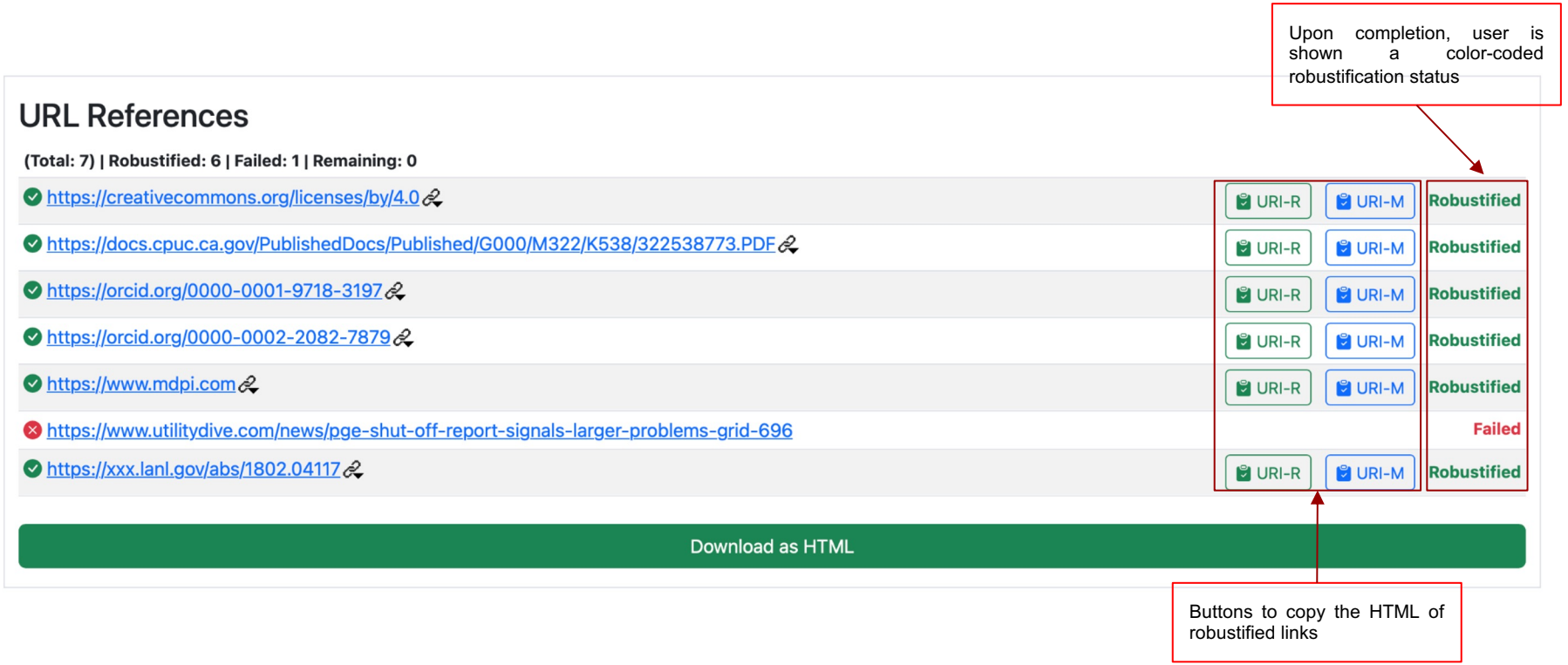
User is shown a spinner to indicate ongoing robustification.

Ongoing Robustification of URLs

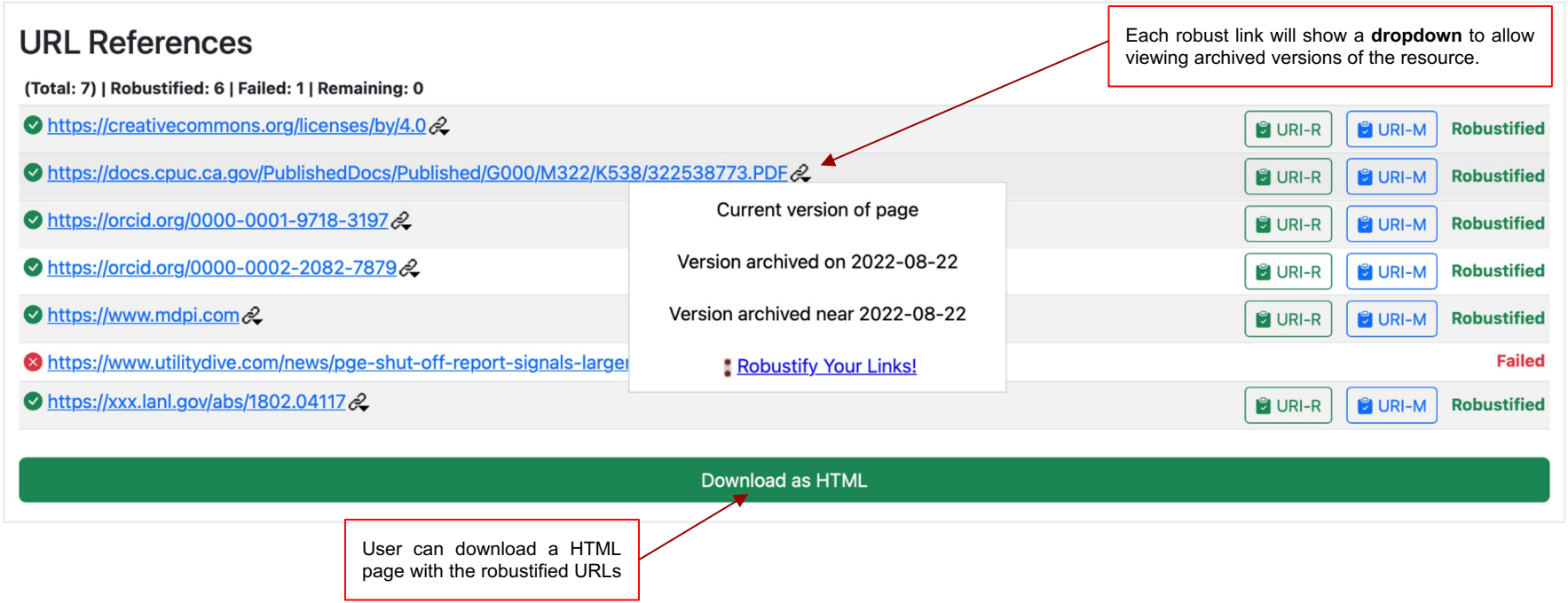
The robustification status is displayed as text



Completed Robustification of URLs



Regular link vs Robust Link



Downloaded Robust Links HTML

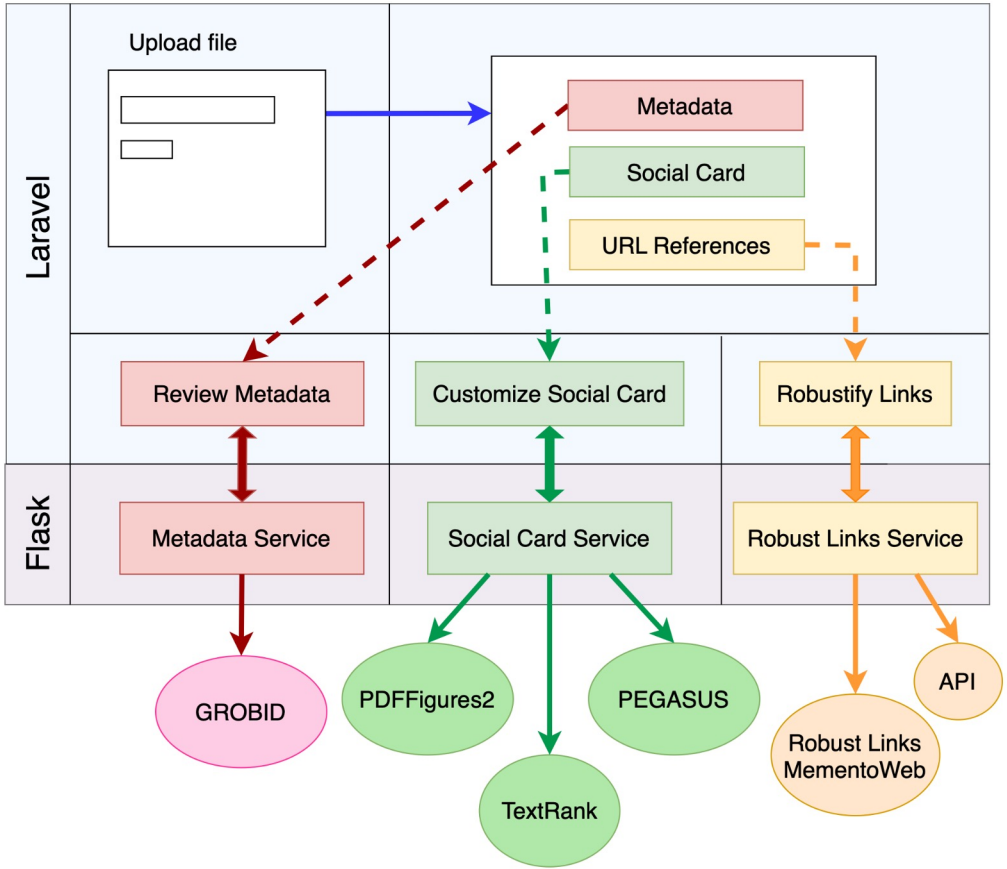
Robust Links | LA-UR-20-29808.pdf

This robust links file was generated by the PDF server project.

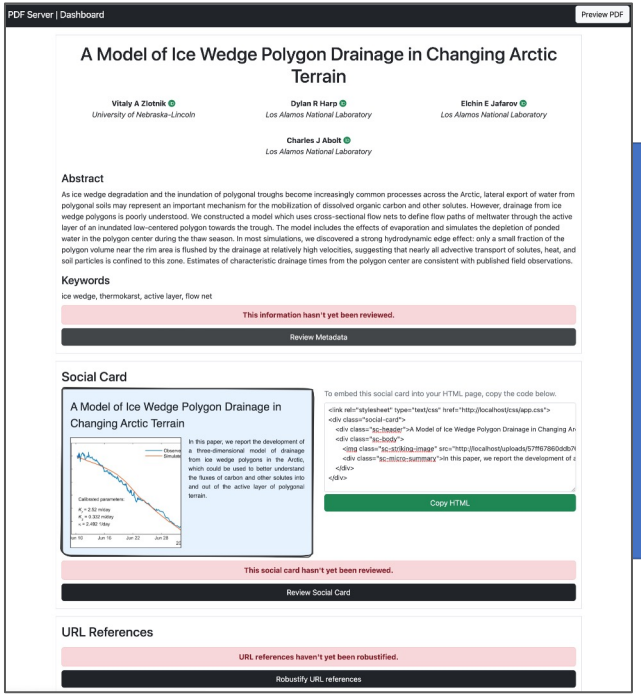
URL Reference	Status
https://creativecommons.org/licenses/by/4.0	robustified
https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M322/K538/322538773.PDF	robustified
https://orcid.org/0000-0001-9718-3197	robustified
https://www.mdpi.com/journal/energies	robustified

This downloaded HTML can be viewed and shared independently.

Demo of PDF Server



Summary



PDF SERVER

Extracting Metadata

Gavindya Jayawardena
gavindya@cs.odu.edu

Generating Social Cards

Himarsha Jayanetti
hjaya002@odu.edu

Robustifying URL References

Yasith Jayawardana
yasith@cs.odu.edu

We thank our supervisors and developers at LANL for their continuous support during our internship.
Supervisors: Martin Klein, Brian Cain, and Shawn M. Jones
Developers: Alex Shocklee, Daniel Waybright, and Robert Kent Hettinga
Teams: ISC and Proto