

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



Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA00001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

PDF Server: "Robustifying" URL References in PDF Documents

Yasith Jayawardana

Proto-team, SRO-RL

Supervisor: Martin Klein



Mini Summer Student Symposium 2022

PhD Student

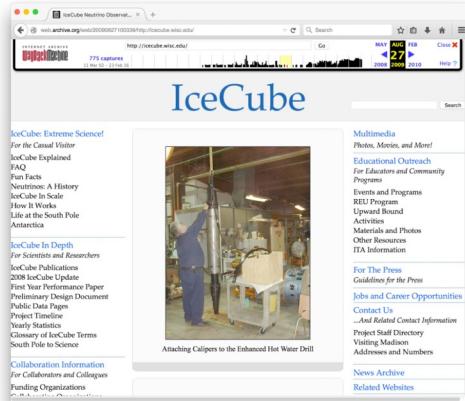
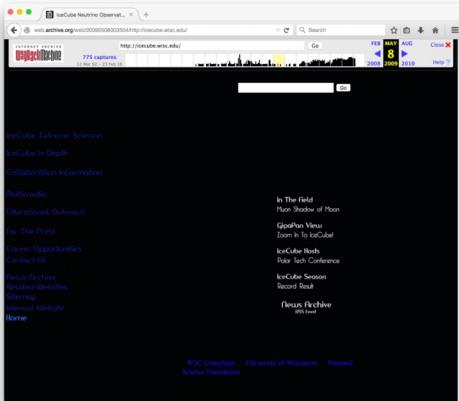
Department of Computer Science, Old Dominion University
Norfolk, Virginia



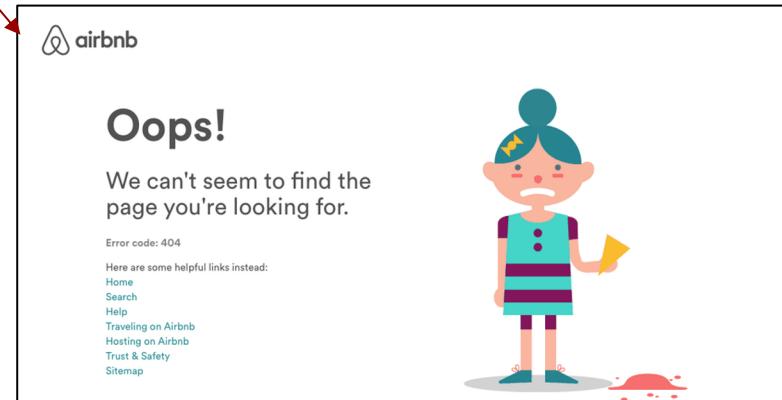
Links on the Web Break all the Time

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

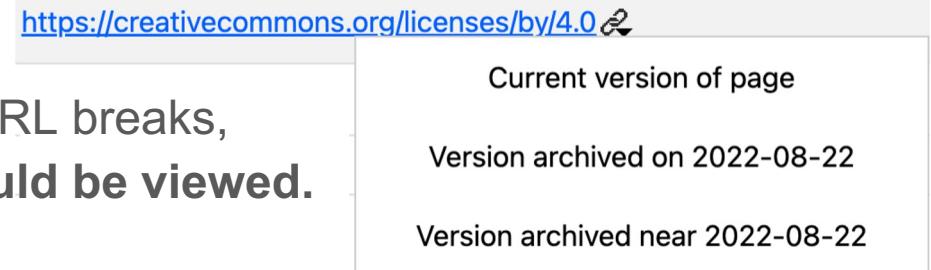
Given an HTML hyperlink,

```
<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>
```

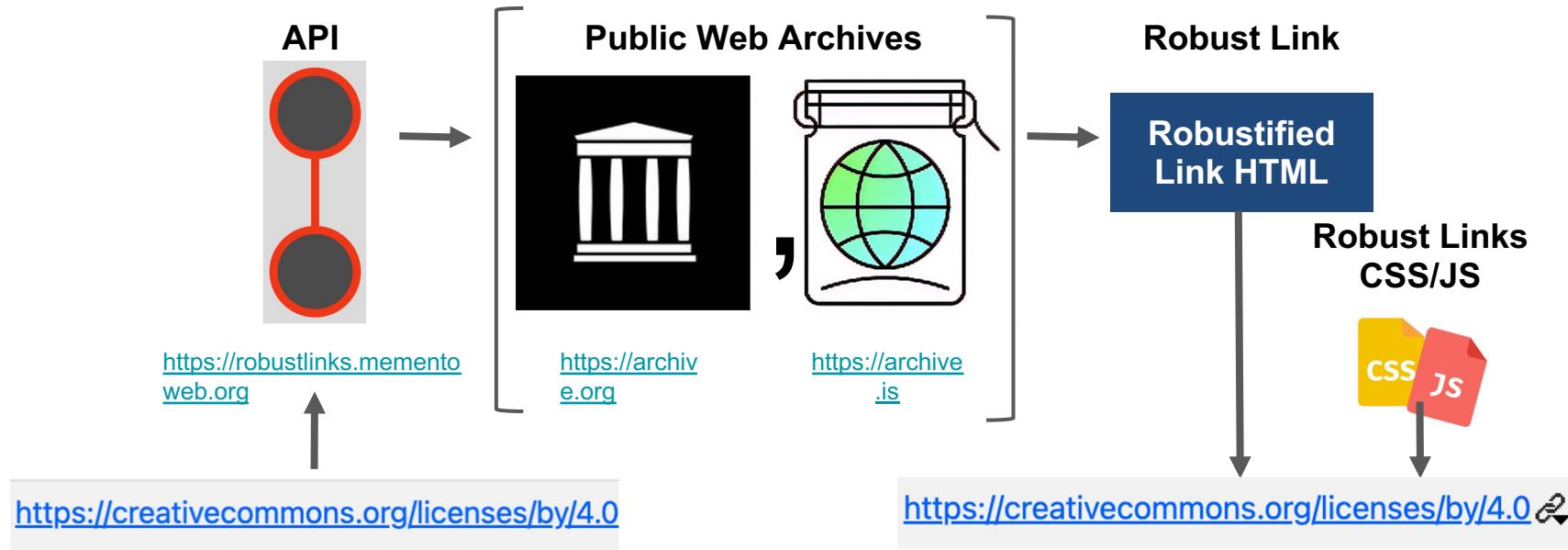


The screenshot shows a browser window with a link to <https://creativecommons.org/licenses/by/4.0>. Below the link, there are three options:

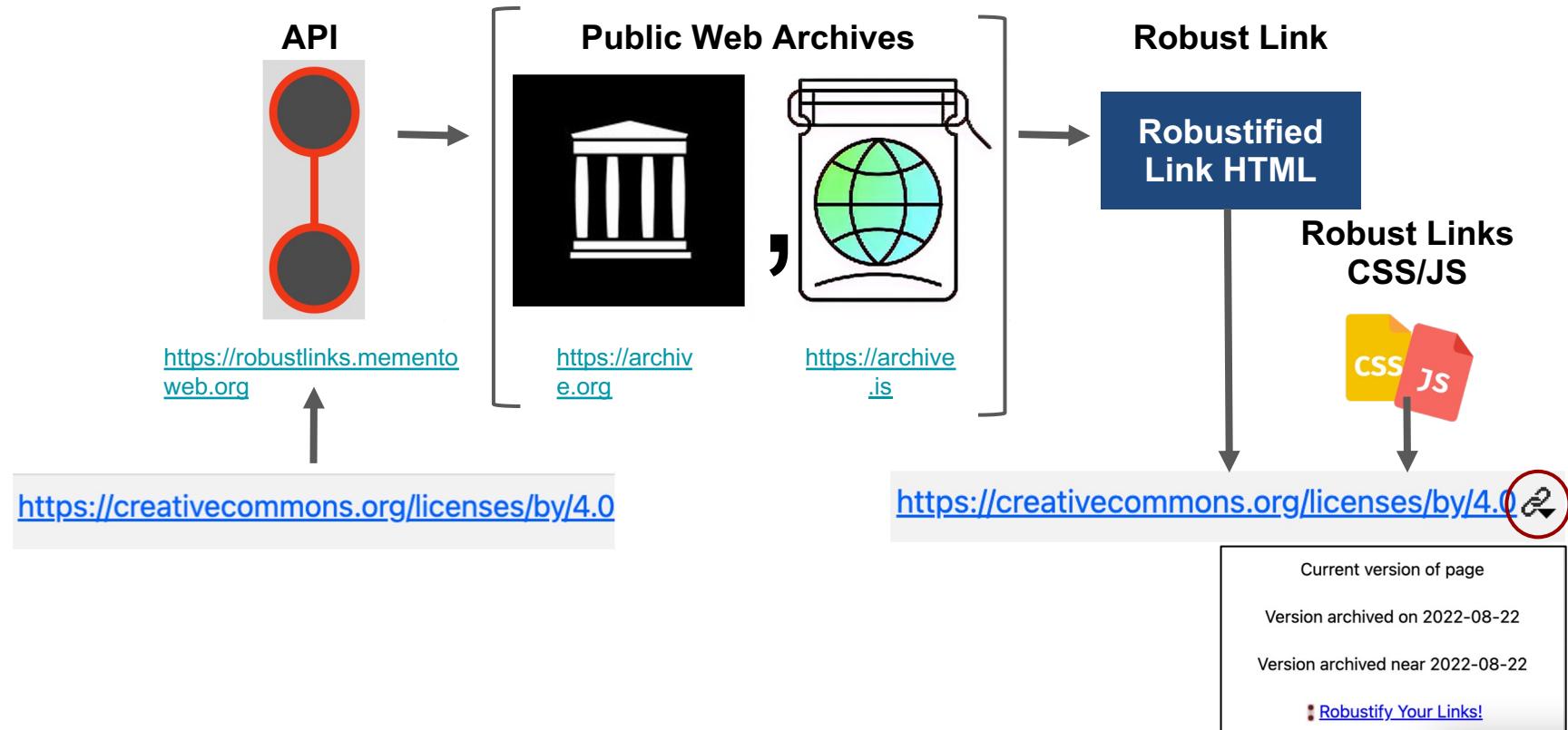
- Current version of page
- Version archived on 2022-08-22
- Version archived near 2022-08-22

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

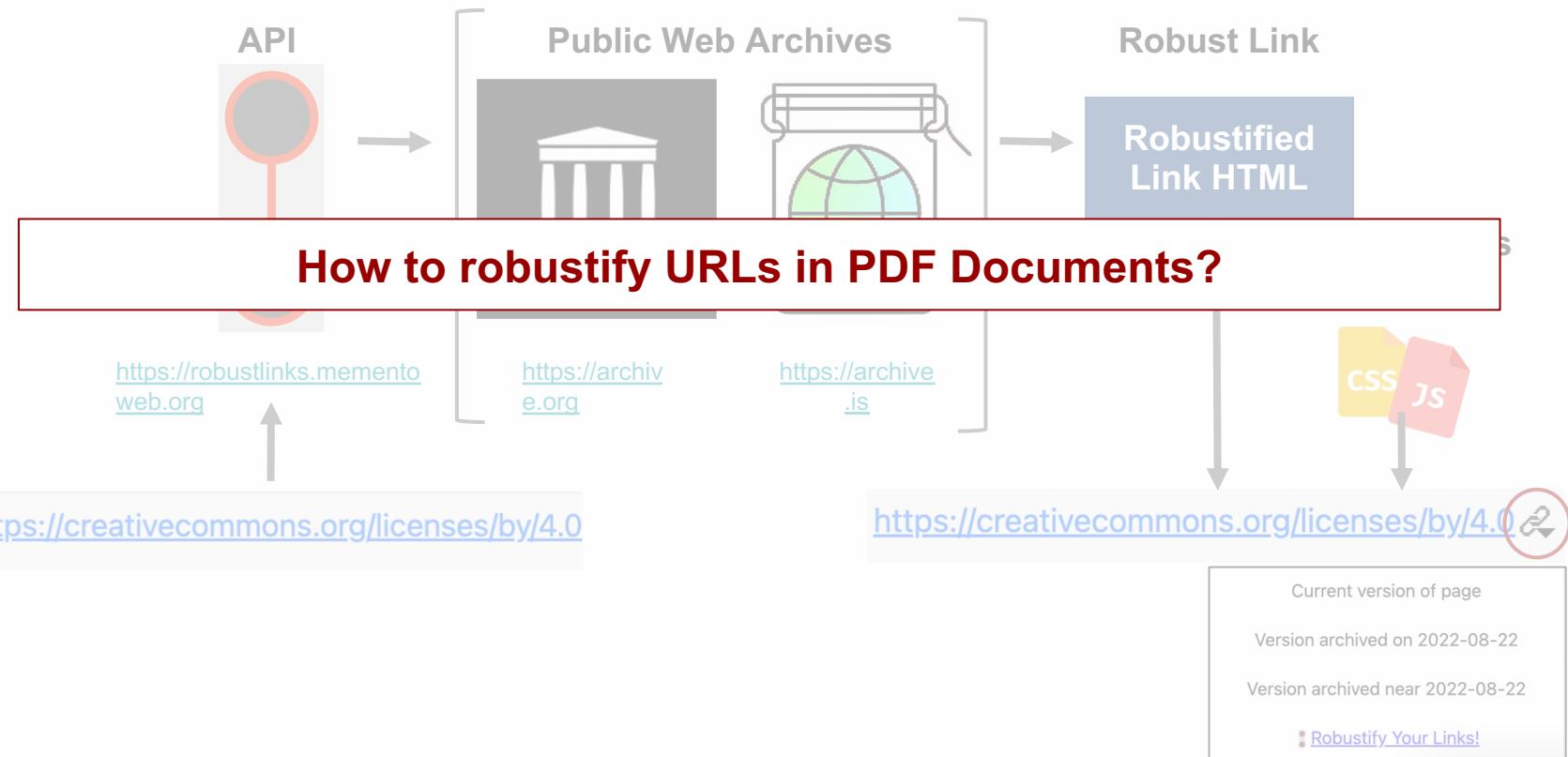
Robustifying URLs using the Robust Links Service



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**

**HTTP has undergone
of organic growth,
implementation based
small devices.**

64 www.computer.org/internet

Authorized licensed use limited to: Old Dominion University. Downloaded on 12/15/2017 10:30 AM

beyond a simple symbiotic relationship to a single gateway node giving them the power of the net is worthwhile.

Another cluster of embedded devices sports around 50 Kbytes of RAM and maybe 250 Kbytes of code space. These *class-2* devices can indeed speak the exact same protocols used among desktops, laptops, and rack-mount servers. However, even these devices can benefit from constrained protocols — they'd use less power and fewer network resources, would leave more functionality available to applications, and could also more easily communicate with *class-1* devices in their environment.

In short, an Internet of Things that wants to make good use of inexpensive class-1 devices, and of

Annotated URL

organic growth, leading to considerable implementation baggage that overwhelms small devices.

However, HTTP is designed to interoperate through proxies; what we really need in constrained environments is REST, not necessarily all HTTP's bells and whistles. CoAP is a fresh approach to a Web application transfer protocol that tries to get by with very limited resources. CoAP isn't just "compressed HTTP" — although it provides the same basic set of services, it does so with a very frugal design (see Figure 2).]

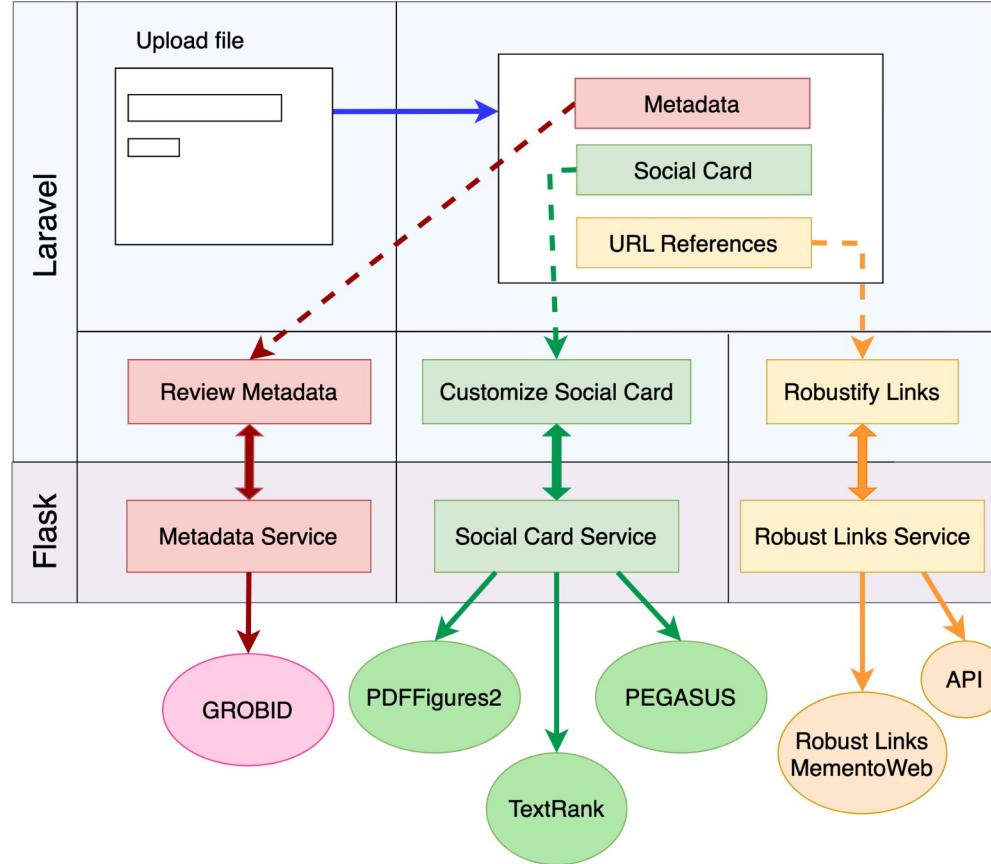
On top of CoAP's message structure, the CoAP base specification defines four request methods: PUT, POST, and DELETE. The response codes are patterned after the HTTP response code structure, with the familiar "404 not found" response, in a single byte ("4.04").

Interworking with HTTP
CoAP would already be useful if we could use it only for translating between CoAP and HTTP, but it reaches its full potential when it can be used to implement a full stack of network protocols.

Non-annotated URL



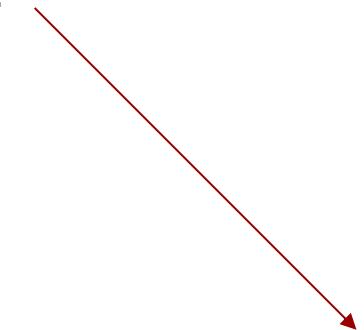
Overall Architecture of PDF Server



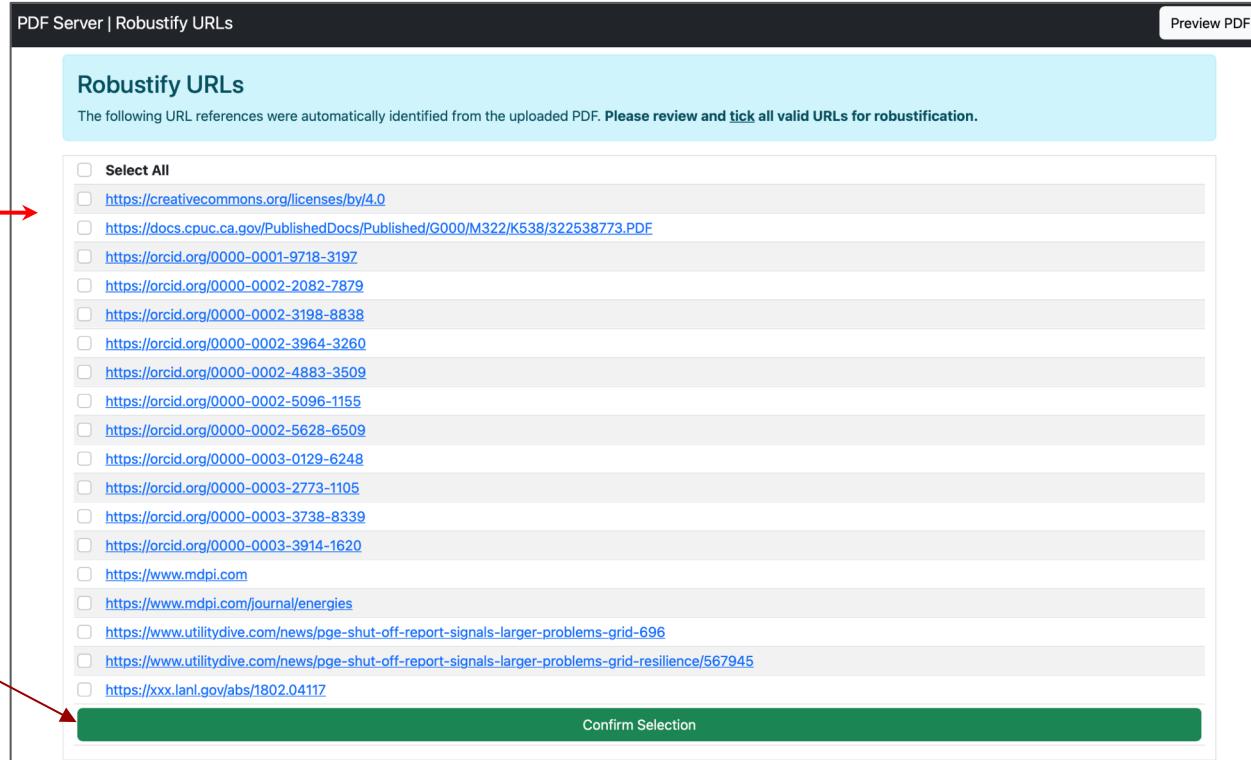
Robustify Links on PDF Server

Idea: 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

PDF Server | Robustify URLs

Robustify URLs

The following URL references were automatically identified from the uploaded PDF. Please review and tick all valid URLs for robustification.

- <https://creativecommons.org/licenses/by/4.0>
- <https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M322/K538/322538773.PDF>
- <https://orcid.org/0000-0001-9718-3197>
- <https://orcid.org/0000-0002-2082-7879>
- <https://orcid.org/0000-0002-3198-8838>
- <https://orcid.org/0000-0002-3964-3260>
- <https://orcid.org/0000-0002-4883-3509>
- <https://orcid.org/0000-0002-5096-1155>
- <https://orcid.org/0000-0002-5628-6509>
- <https://orcid.org/0000-0003-0129-6248>
- <https://orcid.org/0000-0003-2773-1105>
- <https://orcid.org/0000-0003-3738-8339>
- <https://orcid.org/0000-0003-3914-1620>
- <https://www.mdpi.com>
- <https://www.mdpi.com/journal/energies>
- <https://www.utilitydive.com/news/pge-shut-off-report-signals-larger-problems-grid-696>
- <https://www.utilitydive.com/news/pge-shut-off-report-signals-larger-problems-grid-resilience/567945>
- <https://xxx.lanl.gov/abs/1802.04117>

Preview PDF

Confirm Selection

User is shown a spinner to indicate ongoing robustification.

Ongoing Robustification of URLs

URL References

(Total: 7) | Robustified: 0 | Failed: 0 | Remaining: 7

| | |
|---|---------|
| C https://creativecommons.org/licenses/by/4.0 | Pending |
| C https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M322/K538/322538773.PDF | Pending |
| C https://orcid.org/0000-0001-9718-3197 | Pending |
| C https://orcid.org/0000-0002-2082-7879 | Pending |
| C https://www.mdpi.com | Pending |
| C https://www.utilitydive.com/news/pge-shut-off-report-signals-larger-problems-grid-696 | Pending |
| C https://xxx.lanl.gov/abs/1802.04117 | Pending |

The robustification status is displayed as text

Completed Robustification of URLs

URL References

(Total: 7) | Robustified: 6 | Failed: 1 | Remaining: 0

- ✓ <https://creativecommons.org/licenses/by/4.0>
- ✓ <https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M322/K538/322538773.PDF>
- ✓ <https://orcid.org/0000-0001-9718-3197>
- ✓ <https://orcid.org/0000-0002-2082-7879>
- ✓ <https://www.mdpi.com>
- ✗ <https://www.utilitydive.com/news/pge-shut-off-report-signals-larger-problems-grid-696>
- ✓ <https://xxx.lanl.gov/abs/1802.04117>

Download as HTML

| | | |
|---|---|-------------|
| <input checked="" type="checkbox"/> URI-R | <input checked="" type="checkbox"/> URI-M | Robustified |
| <input checked="" type="checkbox"/> URI-R | <input checked="" type="checkbox"/> URI-M | Robustified |
| <input checked="" type="checkbox"/> URI-R | <input checked="" type="checkbox"/> URI-M | Robustified |
| <input checked="" type="checkbox"/> URI-R | <input checked="" type="checkbox"/> URI-M | Robustified |
| <input checked="" type="checkbox"/> URI-R | <input checked="" type="checkbox"/> URI-M | Robustified |
| | | Failed |
| <input checked="" type="checkbox"/> URI-R | <input checked="" type="checkbox"/> URI-M | Robustified |

Upon completion, user is shown a color-coded robustification status

Buttons to copy the HTML of robustified links

Regular link vs Robust Link

URL References

(Total: 7) | Robustified: 6 | Failed: 1 | Remaining: 0

✓ <https://creativecommons.org/licenses/by/4.0> ↗

✓ <https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M322/K538/322538773.PDF> ↗

✓ <https://orcid.org/0000-0001-9718-3197> ↗

✓ <https://orcid.org/0000-0002-2082-7879> ↗

✓ <https://www.mdpi.com> ↗

✗ <https://www.utilitydive.com/news/pge-shut-off-report-signals-large>

✓ <https://xxx.lanl.gov/abs/1802.04117> ↗

Current version of page

Version archived on 2022-08-22

Version archived near 2022-08-22

Robustify Your Links!

URI-R URI-M Robustified

Failed

URI-R URI-M Robustified

Download as HTML

User can download a HTML page with the robustified URLs

Each robust link will show a **dropdown** to allow viewing archived versions of the resource.

Downloaded Robust Links HTML

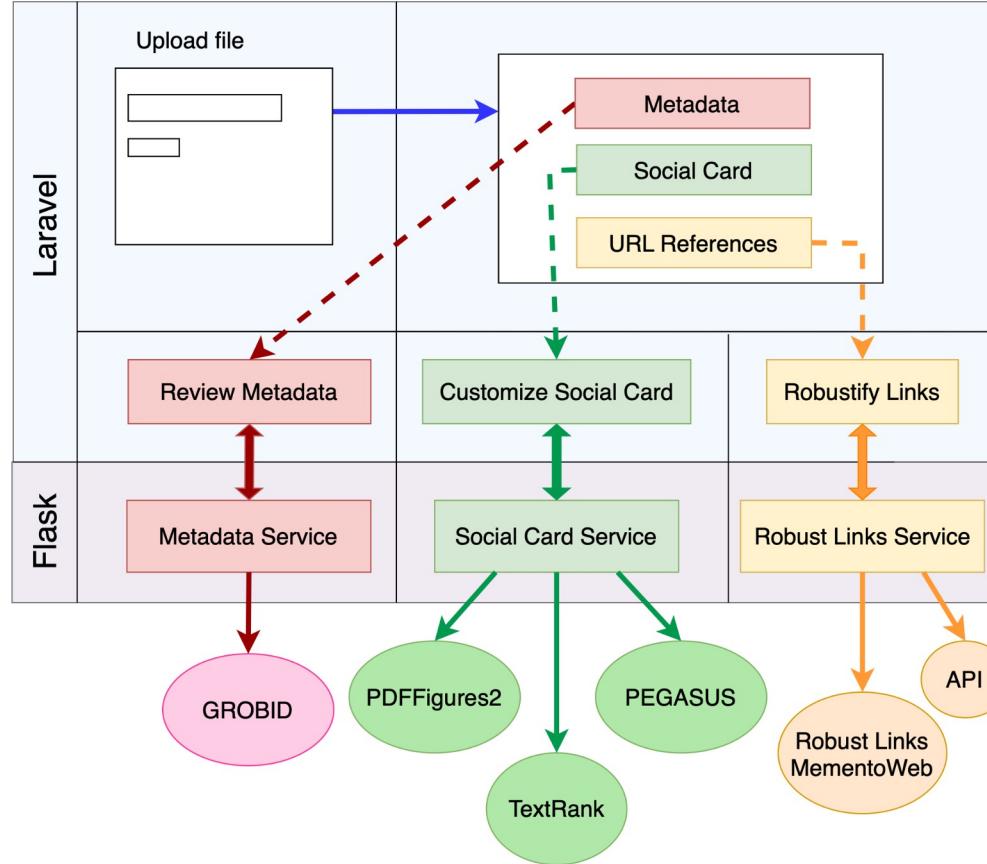
Robust Links | LA-UR-20-29808.pdf

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 | Dashboard

Preview PDF

A Model of Ice Wedge Polygon Drainage in Changing Arctic Terrain

Vitaly A Zlotnik University of Nebraska-Lincoln
Dylan R Harp Los Alamos National Laboratory
Elchin E Jafarov Los Alamos National Laboratory
Charles J Abolt Los Alamos National Laboratory

Abstract

As ice wedge degradation and the inundation of polygonal troughs become increasingly common processes across the Arctic, lateral export of water from polygonal troughs may represent an important mechanism for the mobilization of dissolved carbonic carbon and other solutes. However, drainage from ice wedge polygons is poorly understood. We constructed a model which uses cross-sectional flow nets to define flow paths of meltwater through the active layer of an inundated low-centered polygon towards the trough. The model includes the effects of evaporation and simulates the depletion of ponded water in the polygon center during the thaw season. In most simulations, we discovered a strong hydrodynamic edge effect: only a small fraction of the polygon volume near the rim area is flushed by the drainage at relatively high velocities, suggesting that nearly all advective transport of solutes, heat, and soil particles is confined to this zone. Estimates of characteristic drainage times from the polygon center are consistent with published field observations.

Keywords

ice wedge, thermokarst, active layer, flow net

This information hasn't yet been reviewed.

Review Metadata

Social Card

A Model of Ice Wedge Polygon Drainage in Changing Arctic Terrain

In this paper, we report the development of a three-dimensional model of drainage from ice wedge polygons in the Arctic. The model includes the effects of evaporation and the fluxes of carbon and other solutes into and out of the active layer of polygonal terrain.

Calculated parameters:
 $K = 3.3 \text{ mm/s}$
 $K_s = 0.30 \text{ mm/s}$
 $L = 2.40 \text{ m/s}$

Copy HTML

This social card hasn't yet been reviewed.

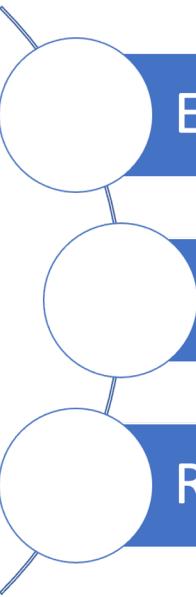
Review Social Card

URL References

URL references haven't yet been robustified.

Robustify URL references

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

