LA-UR-22-29460

Approved for public release; distribution is unlimited.

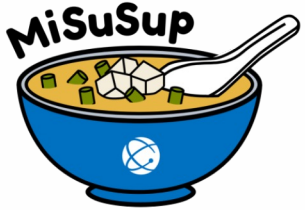
Title:	PDF Server: Extracting Metadata from Scientific PDF Documents
Author(s):	Jayawardena, Gavindya Cain, Brian J. Klein, Martin Jones, Shawn Morgan
Intended for:	Mini Summer Student Symposium 2022, 2022-08-25 (Los Alamos, New Mexico, United States)
Issued:	2022-09-12

Los Alamos
NATIONAL LABORATORY

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 89233218CNA000001. 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: Extracting Metadata from Scientific PDF Documents

Gavindya Jayawardena
Institutional Scientific Content Team, SRO-RL
Supervisor: Brian Cain
Co-Supervisors: Martin Klein, Shawn M. Jones



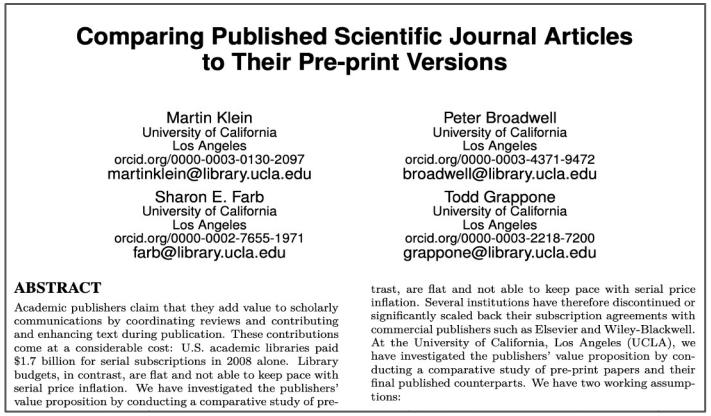
Mini Summer Student Symposium 2022

PhD Student
Department of Computer Science, Old Dominion University
Norfolk, Virginia
@Gavindya2

Los Alamos NATIONAL LABORATORY

WSDL ODU

Extract Metadata from Scientific PDF Documents



Comparing Published Scientific Journal Articles to Their Pre-print Versions

Martin Klein
University of California
Los Angeles
orcid.org/0000-0003-0130-2097
martinklein@library.ucla.edu

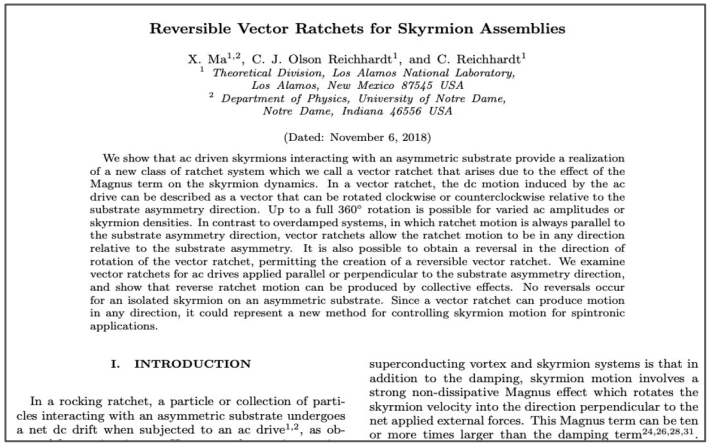
Peter Broadwell
University of California
Los Angeles
orcid.org/0000-0003-4371-9472
broadwell@library.ucla.edu

Sharon E. Farb
University of California
Los Angeles
orcid.org/0000-0002-7655-1971
farb@library.ucla.edu

Todd Grappone
University of California
Los Angeles
orcid.org/0000-0003-2218-7200
grappone@library.ucla.edu

ABSTRACT

Academic publishers claim that they add value to scholarly communications by coordinating reviews and contributing and enhancing text during publication. These contributions come at a considerable cost: U.S. academic libraries paid $1.7 billion for serial subscriptions in 2008 alone. Library budgets, in contrast, are flat and not able to keep pace with serial price inflation. We have investigated the publishers' value proposition by conducting a comparative study of pre-trast, are flat and not able to keep pace with serial price inflation. Several institutions have therefore discontinued or significantly scaled back their subscription agreements with commercial publishers such as Elsevier and Wiley-Blackwell. At the University of California, Los Angeles (UCLA), we have investigated the publishers' value proposition by conducting a comparative study of pre-print papers and their final published counterparts. We have two working assumptions:



Reversible Vector Ratchets for Skyrmion Assemblies

X. Ma[1,2], C. J. Olson Reichhardt[1], and C. Reichhardt[1]
[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545 USA*
[2] *Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556 USA*

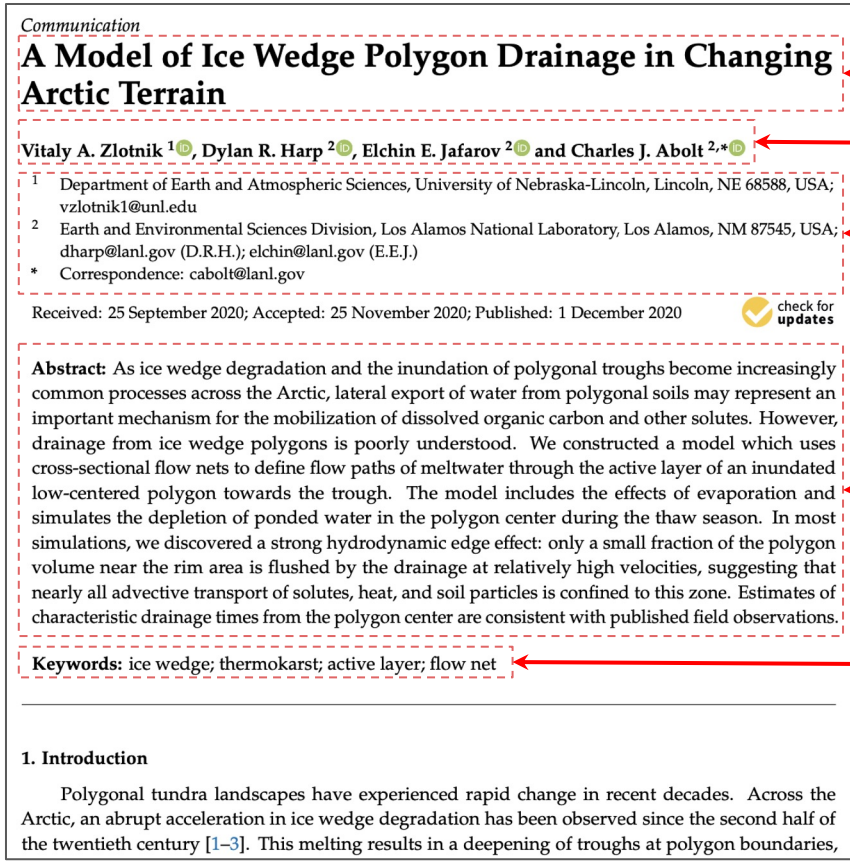
(Dated: November 6, 2018)

We show that ac driven skyrmions interacting with an asymmetric substrate provide a realization of a new class of ratchet system which we call a vector ratchet that arises due to the effect of the Magnus term on the skyrmion dynamics. In a vector ratchet, the dc motion induced by the ac drive can be described as a vector that can be rotated clockwise or counterclockwise relative to the substrate asymmetry direction. Up to a full 360° rotation is possible for varied ac amplitudes or skyrmion densities. In contrast to overdamped systems, in which ratchet motion is always parallel to the substrate asymmetry direction, vector ratchets allow the ratchet motion to be in any direction relative to the substrate asymmetry. It is also possible to obtain a reversal in the direction of rotation of the vector ratchet, permitting the creation of a reversible vector ratchet. We examine vector ratchets for ac drives applied parallel or perpendicular to the substrate asymmetry direction, and show that reverse ratchet motion can be produced by collective effects. No reversals occur for an isolated skyrmion on an asymmetric substrate. Since a vector ratchet can produce motion in any direction, it could represent a new method for controlling skyrmion motion for spintronic applications.

I. INTRODUCTION

In a rocking ratchet, a particle or collection of particles interacting with an asymmetric substrate undergoes a net dc drift when subjected to an ac drive[1,2], as ob-

superconducting vortex and skyrmion systems is that in addition to the damping, skyrmion motion involves a strong non-dissipative Magnus effect which rotates the skyrmion velocity into the direction perpendicular to the net applied external forces. This Magnus term can be ten or more times larger than the damping term[24,26,28,31]



Communication

A Model of Ice Wedge Polygon Drainage in Changing Arctic Terrain

Vitaly A. Zlotnik [1], Dylan R. Harp [2], Elchin E. Jafarov [2] and Charles J. Abolt [2,*]

[1] Department of Earth and Atmospheric Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, USA; vzlotnik1@unl.edu
[2] Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; dharp@lanl.gov (D.R.H.); elchin@lanl.gov (E.E.J.)
* Correspondence: cabolt@lanl.gov

Received: 25 September 2020; Accepted: 25 November 2020; Published: 1 December 2020

check for updates

Abstract: As ice wedge degradation and the inundation of polygonal troughs become increasingly common processes across the Arctic, lateral export of water from polygonal soils may represent an important mechanism for the mobilization of dissolved organic 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

1. Introduction

Polygonal tundra landscapes have experienced rapid change in recent decades. Across the Arctic, an abrupt acceleration in ice wedge degradation has been observed since the second half of the twentieth century [1–3]. This melting results in a deepening of troughs at polygon boundaries,

Title

Authors

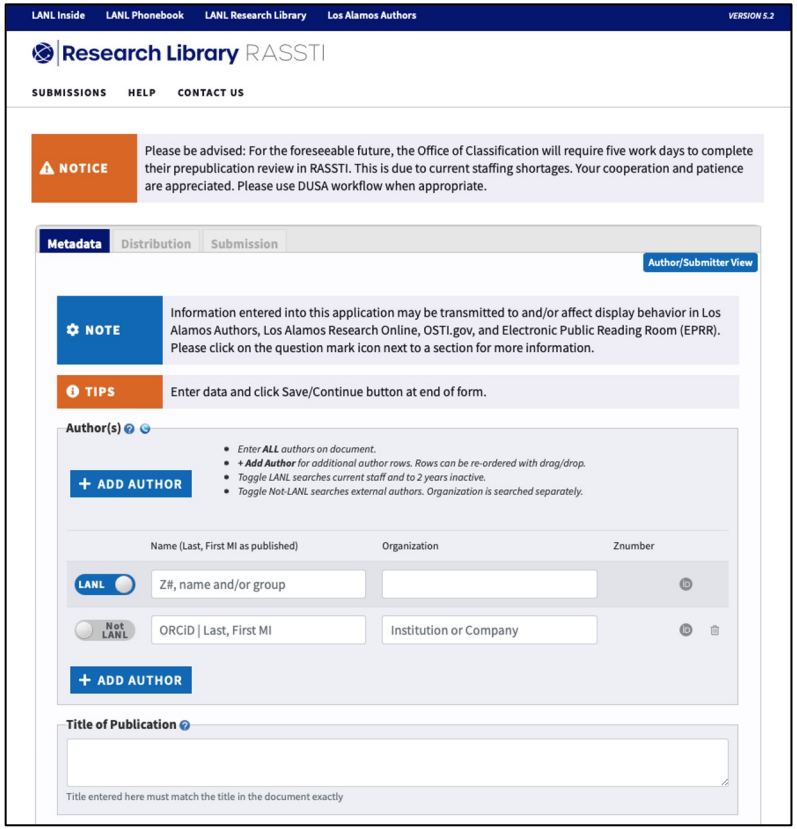
Affiliations

Abstract

Keywords

Ingest Extracted Metadata into RASSTI

- RASSTI is the starting point for collecting LANL's documented scientific output
- It provides workflows for authors and reviewers
- Input:
  - Author entered metadata
  - Associated PDF
- Output:
  - Approvals from reviewers containing Los Alamos Report Number



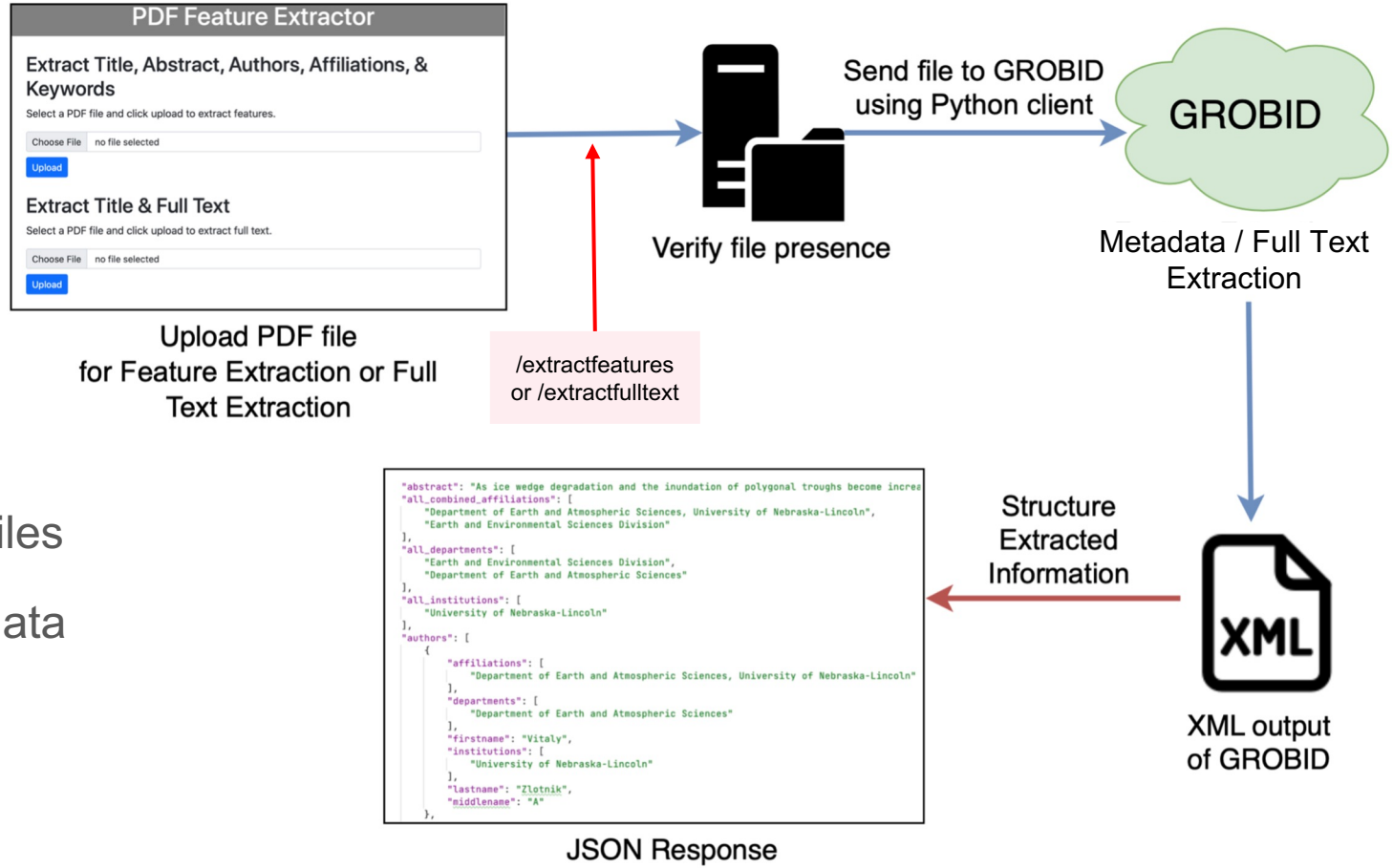
RASSTI Review & Approval System

In Summer 2021, the Extraction of Metadata from Scientific PDF Documents in Real-time was completed using GROBID (GeneRation Of Bibliographic Data)

- **GROBID:** Machine learning library for extracting information from scholarly PDF documents
- **Main functions of GROBID:**
  - Metadata extraction (title, abstract, authors, affiliations, keywords)
  - Full text extraction (paragraphs, section titles)

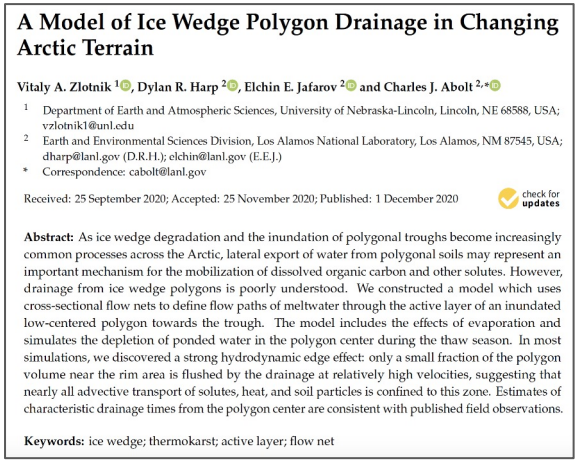
System Overview

- **Flask** web application
- Upload PDF files
- Extract Metadata or Full Text



Reviewing Extracted Metadata is Essential

- Machine-Learning systems are not always 100% accurate
- Tested the usability of GROBID when a PDF is submitted
  - It gave an accuracy **> 75%** for all features except affiliations
  - On average, it extracted **> 60%** of actual affiliations
- Results of the evaluation yields a promising usability of GROBID with reviews



A Model of Ice Wedge Polygon Drainage in Changing Arctic Terrain

Vitaly A. Zlotnik [1], Dylan R. Harp [2], Elchin E. Jafarov [2] and Charles J. Abolt [2,*]

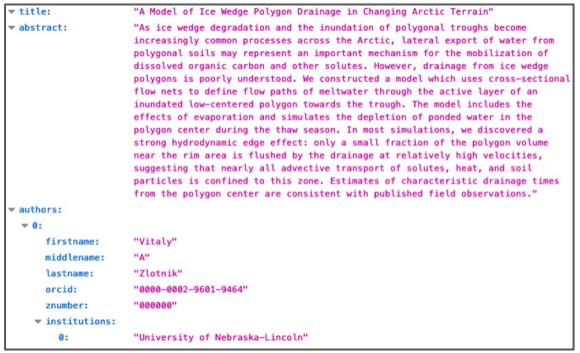
[1] Department of Earth and Atmospheric Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, USA; vzlotnik1@unl.edu
[2] Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; dharp@lanl.gov (D.R.H.); elchin@lanl.gov (E.E.J.)
* Correspondence: cabolt@lanl.gov

Received: 25 September 2020; Accepted: 25 November 2020; Published: 1 December 2020

check for updates

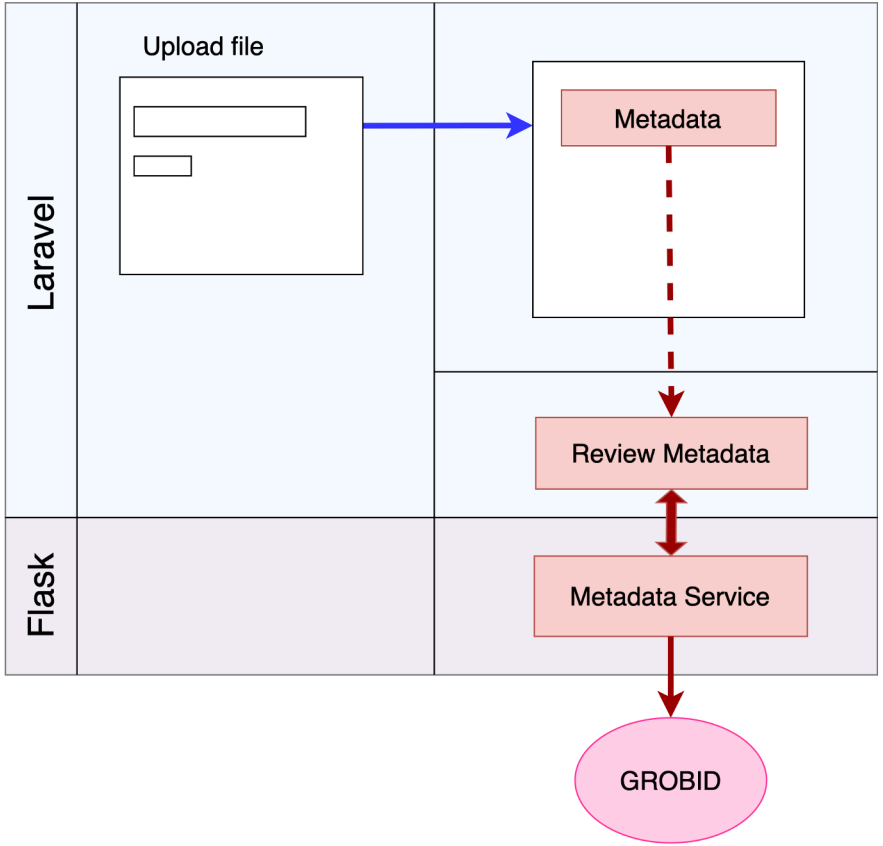
Abstract: As ice wedge degradation and the inundation of polygonal troughs become increasingly common processes across the Arctic, lateral export of water from polygonal soils may represent an important mechanism for the mobilization of dissolved organic carbon and other solutes. However, drainage from 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

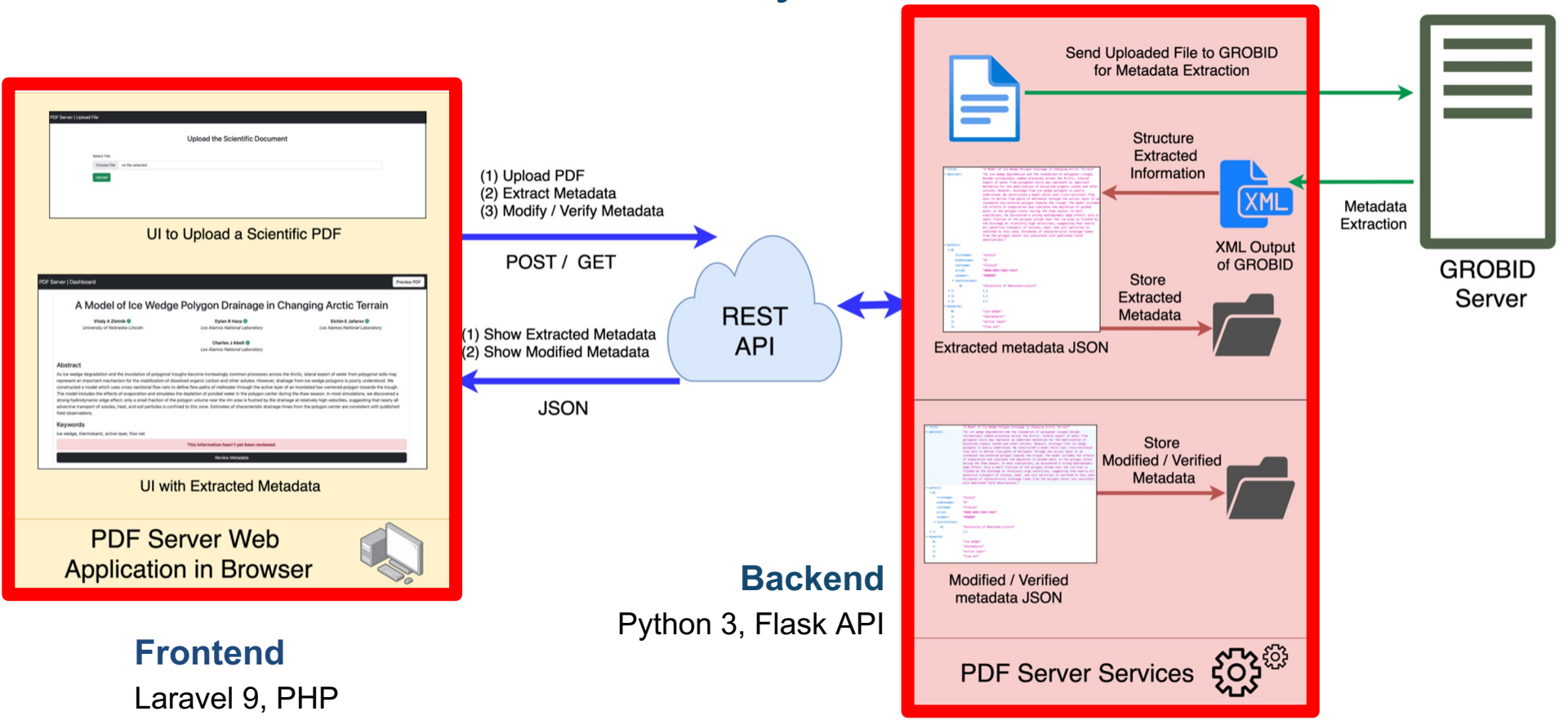



```
title: "A Model of Ice Wedge Polygon Drainage in Changing Arctic Terrain"
abstract: "As ice wedge degradation and the inundation of polygonal troughs become
  increasingly common processes across the Arctic, lateral export of water from
  polygonal soils may represent an important mechanism for the mobilization of
  dissolved organic 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."
authors:
  0:
    firstname: "Vitaly"
    middlename: "A"
    lastname: "Zlotnik"
    orcid: "0000-0002-9601-9464"
    znumber: "000000"
    institutions:
      0: "University of Nebraska-Lincoln"
```

This Summer, the Goal is to Build a User Interface to Review the Extracted Metadata as a component of the PDF Server



Overview of Metadata System within PDF Server



Frontend

Laravel 9, PHP

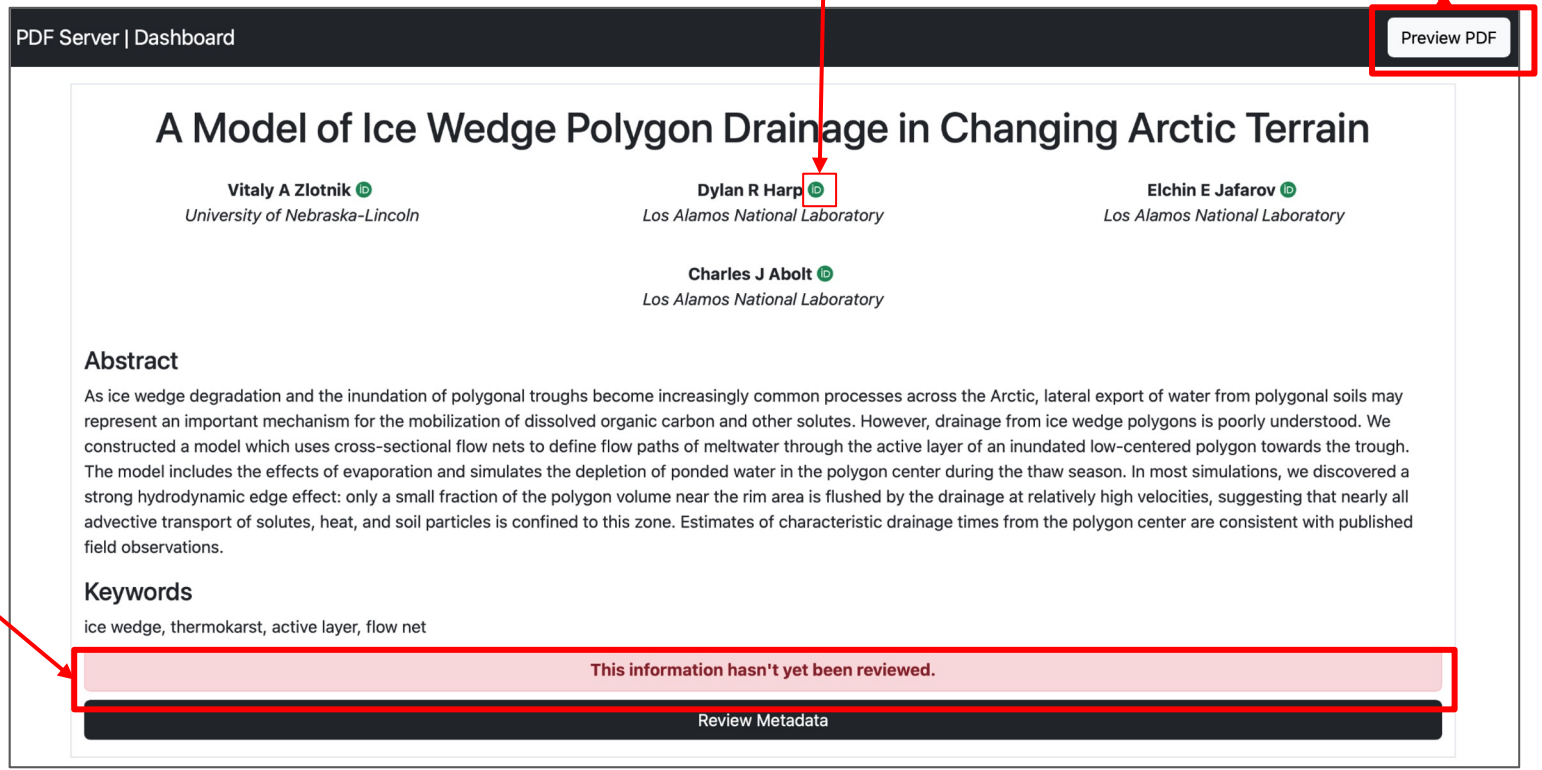
Backend

Python 3, Flask API

Extracted Metadata

ORCID link of the author

Button to view the Uploaded PDF

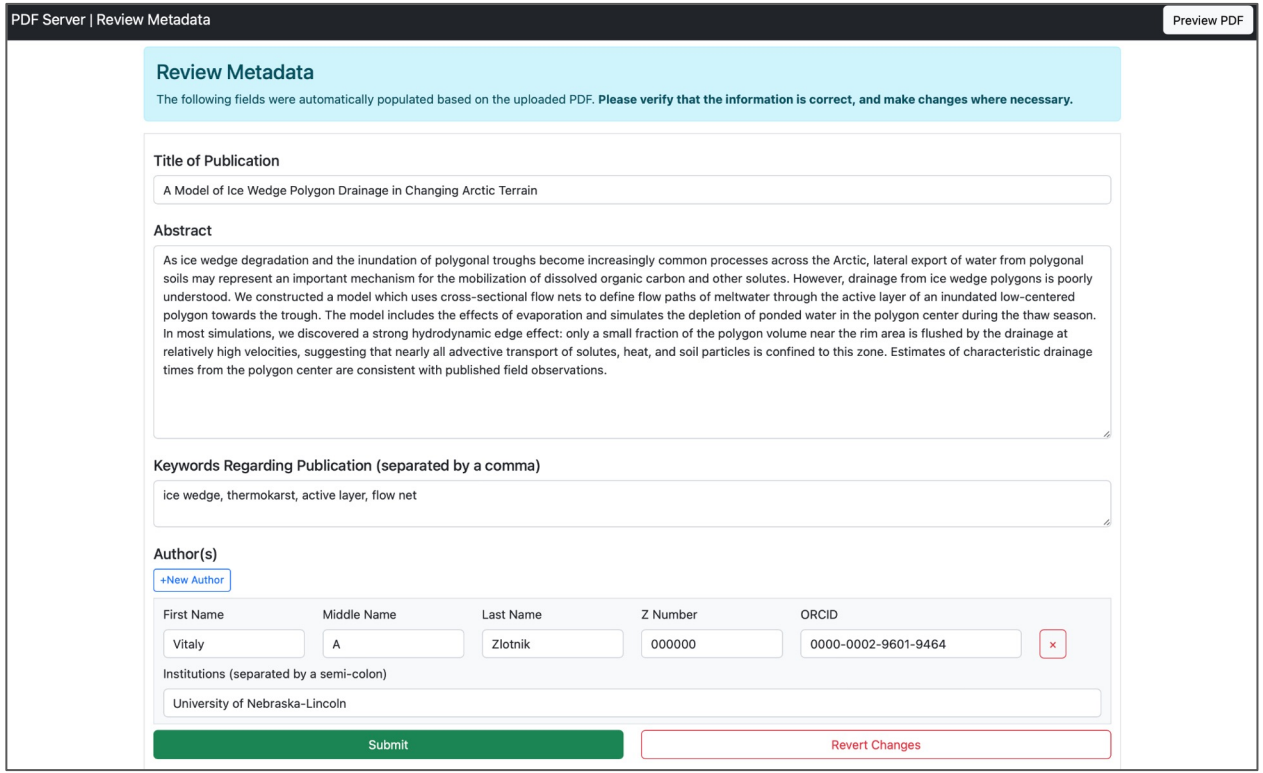


Banner indicating this page is loaded with automatically extracted data

Review Metadata

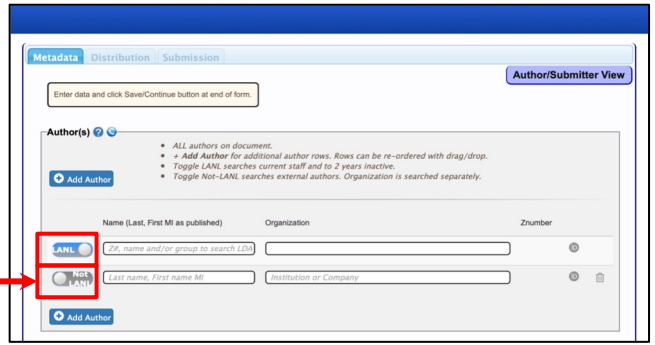
This UI allows users to:

- Review automatically populated metadata based on the uploaded PDF
- Modify,
  - Title
  - Abstract
  - Keywords
  - Authors (First name, Middle name, Last name, ORCID, Znumber if a LANL Employee, Institutions)
    - Delete Authors
    - Add new Authors
- Revert the changes to the automatically extracted metadata



Future Work

- Distinguish between internal and external author
- Overcome GROBID imperfections
  - ORCID is extracted as postcode
  - Subscripts/ Superscripts in texts are not extracted properly




```
<persName><forename type="first">Martin</forename><surname>Klein</surname></persName>
<email>martinklein@library.ucla.edu</email>
<affiliation key="aff0">
<note type="raw_affiliation">University of California Los Angeles orcid.org/0000-0003-0130-2097</note>
<orgName type="institution">University of California Los Angeles orcid</orgName>
<address>
<postCode>0000-0003-0130-2097</postCode>
</address>
</affiliation>
</author> s
```


Contact information:
gavindya@lanl.gov, gavindya@cs.odu.edu
Twitter: @Gavindya2