

Outline

1. Archiving (Digital) Scholarship
2. Conceptuals / Technicals
3. Tools & Crawls
4. Le Catalog
5. IA Scholar

goal goal

- IA open infrastructure for public good/access
- Instead of specialized archiving services...
 - **Add “open scholarship” automation to the existing automation/scale of web harvesting**
- Instead of specialized curation and ingest...
 - **Develop tools to identify scholarly objects in existing harvests/archives, ID/correct incompleteness, augment with metadata**

Perpetual Access Archiving Challenges

- **Print >>> Digital = custodial challenges**
- **Traditional curation doesn't scale**
- **Traditional deposit/services obsolete**
- **First-world preservation non-problems**
- **Long Tail: not English, STEM, or US/EU**
- **Even more problems for web/born-digital OA**

Access

Perpetual Access Access Challenges

- **Wayback Machine = known URLs**
 - **But has ~800M PDFs (~5% scholarly)**
- **Web archives not (yet) in most search**
- **Scholarly outputs all over the dang web**
- **No QA/QC mechanisms or edit tools**
- **(Prior) no targeted, integrated harvesting**



Methods / Approaches

Top-Down, Known-Work Approach

Harvest/archive PIDs, registries, metadata, manifests, etc to find web-published scholarly outputs to archive, extract/augment with metadata, etc

200+ million URLs processed

Methods / Approaches

Top-Down, Domain-Level Approach

Spider Journal, Publisher,
Platform homepages

Cheap + Simple Heritrix Crawls
30+ TB new content per annum



Methods / Approaches

Bottom-Up Approach

Use machine learning tools to identify scholarly objects in TB/PB scale web archives, assess completeness, and match with versions, metadata, etc

~800+ million PDFs in Wayback

Methods / Approaches

Forever Approach

- Partnerships, services
- Open APIs, code, infrastructure
- Add to discovery services
- Add to data services
- Re-distribution & bulk access



Methods / Approaches

Web Crawl Challenges

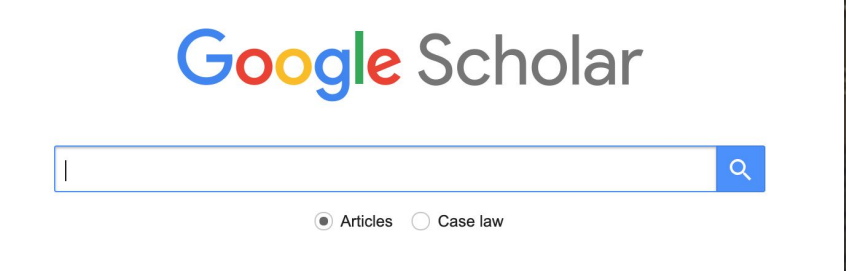
- Javascript-only platforms
- Bot blocking, rate limiting
- Longtail, historical, and bespoke landing page designs
- Journal homepage URL link rot



PARTNER



PARTNER



Google Scholar

Search bar with a magnifying glass icon

Articles Case law



Te Puna Mātauranga o Aotearoa

NATIONAL LIBRARY
OF NEW ZEALAND



nl

Leabharlann Náisiúnta na hÉireann
National Library of Ireland



K

Keepers Registry



LEAN LIBRARY

A SAGE Publishing Company



INTERNET ARCHIVE

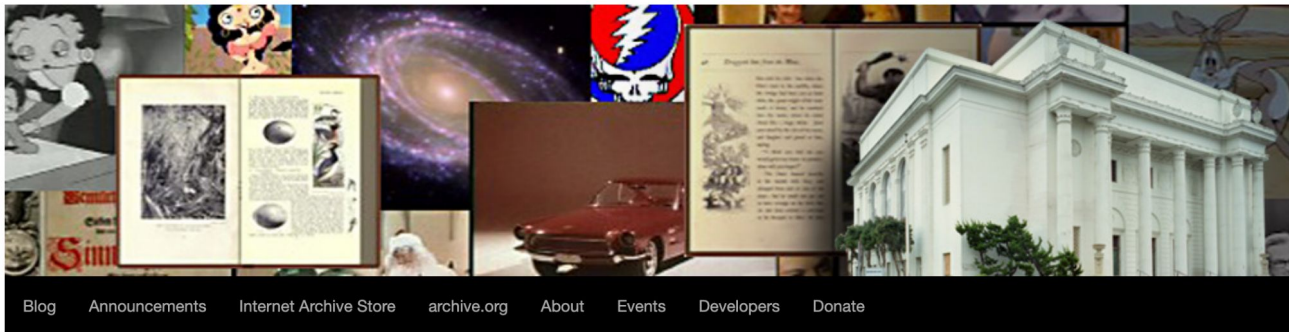


ARCHIVE-IT

PARTNER

Internet Archive Blogs

A blog from the team at archive.org



[Blog](#) [Announcements](#) [Internet Archive Store](#) [archive.org](#) [About](#) [Events](#) [Developers](#) [Donate](#)

Internet Archive Participates in DOAJ-Led Collaboration to Improve the Preservation of OA Journals

Posted on [November 5, 2020](#) by [jefferson](#)

Recent Posts

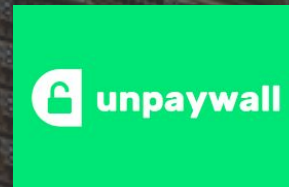


[blog post](#)

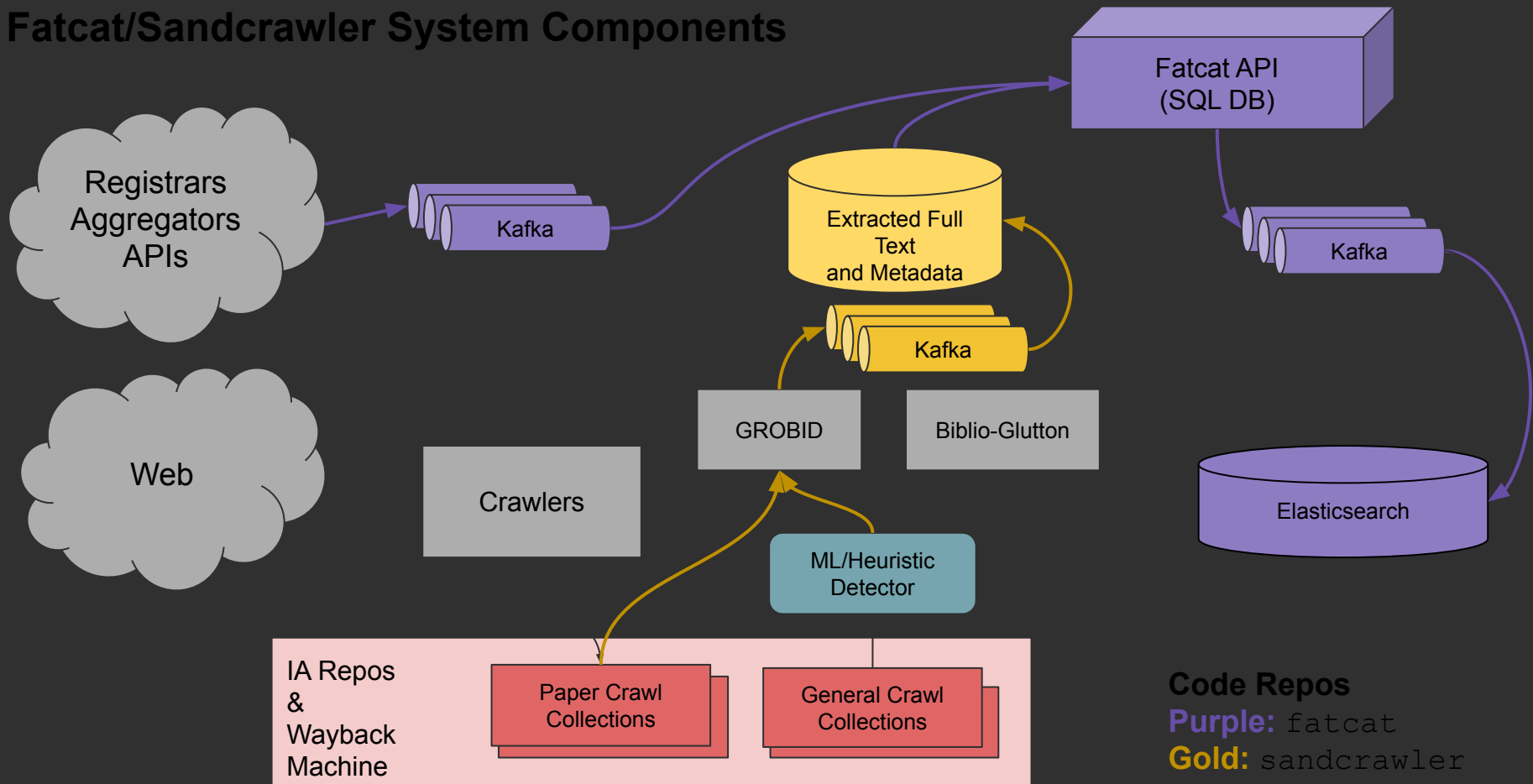
Ingest: Crawl Existing Indices



- Usual suspect (registrars, indices)
 - Transform metadata, import
- Bulk URL seedlist crawling, import “hits”
- Work with peer aggregators hosting fulltext
 - OAI-PMH / IRs



Fatcat/Sandcrawler System Components





Bulk Bibliographic Metadata

Internet Archive Web Group

This collection contains both external ("upstream") metadata dumps and Internet Archive generated databases and reports on our holdings of papers, books, and other documents.

- Share
- Favorite
- RSS
- Edit
- History
- Play All

ABOUT

COLLECTION

https://archive.org/details/ia_biblio_metadata

82 RESULTS

Search this Collection

Metadata

Text contents

PART OF

[The Internet Archive](#)

Media Type

data 81

texts 1

Year

2019 17

2018 22

2017 11

2016 3

2015 1

2014 1

[More](#)

Topics & Subjects Aa

SORT BY VIEWS · TITLE · DATE

Crossref

Crossref DOI Dump (2018-01)

459 1 0

Crossref

Crossref DOI Dump (2018-09)

259 0 0

Microsoft Academic Graph (2016-02-05 snapshot)

by Microsoft Academic Search

233 0 0

oaDOI DOI/URL Dataset

226 0 0

Sci-Hub DOI List

by Sci-Hub

225 0 0

Internet Archive Paper Manifest (2017-09-19)

151 0 0

Internet Archive Paper Manifest (2018-01-25)

136 0 0

CiteSeerX Database Dump (2017-03-31)

by CiteSeerX Group at PSU

128 0 0

Open Academic Graph (aminer.org)

by aminer.org

121 0 0

ORCID Public Data File (2017)

by ORCID, Inc

84 0 0

ROAD/ISSN Directory (2018)

by ROAD: Directory of Open Access Scholarly Resources

84 0 0

DOAJ Journal and Article Metadata (2018)

by Directory of Open Access Journals

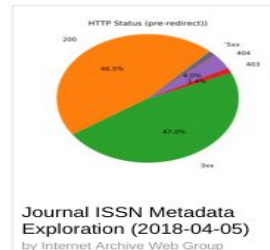
71 0 0

CORE Open Access Paper Metadata (2017-11-)

68 0 0

Internet Archive Paper Manifest (2017-10-06)

65 0 0




Ingest: Index Web Archives


- 1. Petabytes of unsorted web resources**
 - HTML, PDF, Datasets, everything
- 2. Filter down to likely research pubs**
- 3. Try matching against existing catalog**
- 4. Else, create new catalog records**

PDF Tooling



pdf_trio

 Search or jump to... /

 [internetarchive / pdf_trio](https://github.com/internetarchive/pdf_trio)
forked from tralfamadude/pdf_trio

https://github.com/internetarchive/pdf_trio

Making A Production Classifier Ensemble

A ready to use PDF classifier service using BERT, Inception, and fastText

<https://towardsdatascience.com/making-a-production-classifier-ensemble-2d87fbf0f486>

PDF Tooling



pdf_trio



grobid

```
<?xml version="1.0"?>
<quiz>
  <qanda seq="1">
    <question>
      Who was the forty-second
      president of the U.S.A.?
    </question>
    <answer>
      William Jefferson Clinton
    </answer>
  </qanda>
  <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

buzzycat



PG3557
.R5355 Grisham, John
F57 1991

The firm / John Grisham. 1st. ed.
New York : Doubleday, c1991.
431p. ; 24 cm.

1. Government investigators--Fiction.
2. Organized crime--Fiction.



<https://fatcat.wiki/>

Perpetual Access to Millions of Open Research Publications From Around The World

by title, authors, identifiers...

Search

96,947,165
Papers

18,117,429
Fulltext

140,085
Journals



Fatcat is a versioned, user-editable catalog of research publications including journal articles, conference proceedings, and datasets

Features include archival file-level metadata (verified digests and long-term copies), an [open, documented API](#), and work/release indexing (eg, distinguishing between and linking pre-prints, manuscripts, and version-of-record). [Read more...](#)

This service is hosted at [The Internet Archive](#), a US non-profit dedicated to providing Universal Access to All Knowledge. [Donations welcome!](#)

Development funding comes from [The Andrew Mellon Foundation](#) to improve preservation and access to "long-tail" open access works on the public web which might otherwise be lost.



Basic Reader Access

fatcat! About Guide Changelog Search Papers... Login/Signup

Search all Releases

ellen spertus

Can also lookup by identifier or search for containers (eg, journals). Fulltext Available Only

Showing top 23 out of 23 results for eLlen spertus

- ParaSite: mining structural information on the Web**
Ellen Spertus
1997 Computer networks and ISDN systems
doi:10.1016/S0169-7552(97)00033-0
- Dataflow Computation for the J-Machine**
Ellen Spertus
1990
doi:10.21236/ada228612
- Gender benders**
Ellen Spertus
2002 ACM SIGCSE Bulletin
doi:10.1145/543812.543848
- Leveraging an alternative source of computer scientists**
Sheila Humphreys, Ellen Spertus
2002 ACM SIGCSE Bulletin
doi:10.1145/543812.543830
- Squeal: a structured query language for the Web**
Ellen Spertus, Lynn Andrea Stein
2000 Computer Networks
doi:10.1016/S1389-1286(00)00074-8
- Evaluating the locality benefits of active messages**
Ellen Spertus, William J. Dally
1995 Proceedings of the fifth ACM SIGPLAN symposium

Search

INTERNET ARCHIVE waybackmachine <http://people.mills.edu/spertus/Papers/parasite97.pdf> Go SEP 18 APR 18 JUL 18 10 captures 18 Nov 2006 - 18 Aug 2017 2006 2007 2008 About this capture

ParaSite: Mining Structural Information on the Web 1 / 13

http://www.mills.edu/ACAD_INFOMCS/SPERTUS/Parasite/parasite.html

Appearing in *The Sixth International World Wide Web Conference*, April 1997.

ParaSite: Mining Structural Information on the Web

Ellen Spertus
MIT Artificial Intelligence Lab and University of Washington Dept. of CSE
University of Washington
Box 352350
Seattle, WA 98195-2350
elens@ai.mit.edu

Abstract

Web information retrieval tools typically make use of only the text on pages, ignoring valuable information implicitly contained in links. At the other extreme, viewing the Web as a traditional hypertext system would also be a mistake, because heterogeneity, cross-domain links, and the dynamic nature of the Web mean that many assumptions of typical hypertext systems do not apply. The novelty of the Web leads to new problems in information access, and it is necessary to make use of the new kinds of information available, such as multiple independent categorization, naming, and indexing of pages. This paper discusses the varieties of link information (not just hyperlinks) on the Web, how the Web differs from conventional hypertext, and how the links can be exploited to build useful applications. Specific applications presented as part of the ParaSite system find individuals' homepages, new locations of moved pages, and unindexed information.

Introduction

The World-Wide Web contains millions of pages of data. Practical access to this information requires applying and expanding hypertext research to build powerful search tools. Most Web search tools only make use of the text on a page, ignoring another rich source of information, the links among pages. Much human thought has gone into creating each hyperlink and labeling it with anchor text. Other valuable relational information can be gleaned from the structure, hierarchy, and similarity of pieces of text. This information is already used by individuals when they browse the Web. It should be harnessed to build powerful automatic search tools.

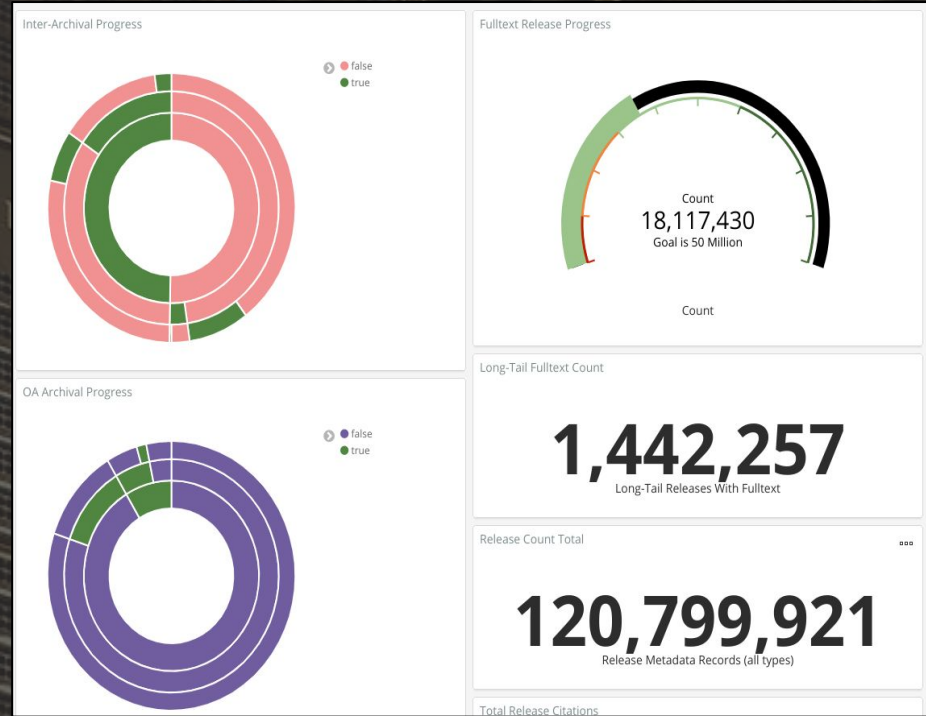
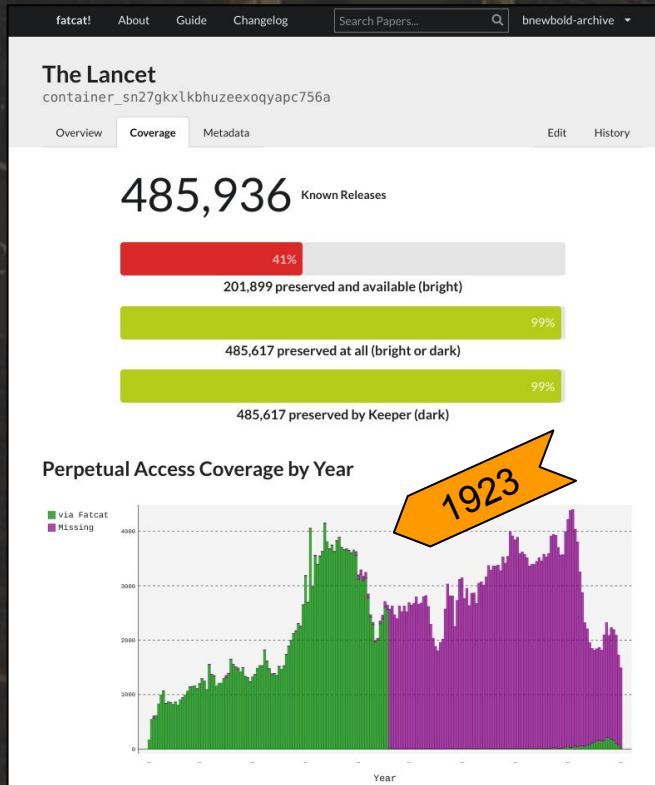
Hypertext research has primarily focused on a single document or set of related documents converted to

Wayback Replay

IIPC 2021



Coverage Dashboards



fatcat! About Guide Changelog Search Papers... bnewbold-archive

The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles

release_hsmo6p4smrganpb3fndaj2lon4

by Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, Stefanie Haustein

Overview Authors (9) References (52) Metadata Edit History

Abstract

<jats:p>Despite growing interest in Open Access (OA) to scholarly literature, there is an unmet need for large-scale, up-to-date, and reproducible studies assessing the prevalence and characteristics of OA. We address this need using oaDOI, an open online service that determines OA status for 67 million articles. We use three samples, each of 100,000 articles, to investigate OA in three populations: (1) all journal articles assigned a Crossref DOI, (2) recent journal articles indexed in Web of Science, and (3) articles viewed by users of Unpaywall, an open-source browser extension that lets users find OA articles using oaDOI. We estimate that at least 28% of the scholarly literature is OA (19M in total) and that this proportion is growing, driven particularly by growth in Gold and Hybrid. The most recent year analyzed (2015) also has the highest percentage of OA (45%). Because of this growth, and the fact that readers disproportionately access newer articles, we find that Unpaywall users encounter OA quite frequently: 47% of articles they view are OA. Notably, the most common mechanism for OA is not Gold, Green, or Hybrid OA, but rather an under-discussed category we dub Bronze: articles made free-to-read on the publisher website, without an explicit Open license. We also examine the citation impact of OA articles, corroborating the so-called open-access citation advantage: accounting for age and discipline, OA articles receive 18% more citations than average, an effect driven primarily by Green and Hybrid OA. We encourage further research using the free oaDOI service, as a way to inform OA policy and practice.</jats:p>

In application/xml+jats format

Published in [PeerJ](#) by PeerJ

Known Files and URLs

application/pdf 2.4 MB sha1:bca1531b0562c6d72e9c...	web.archive.org (webarchive) peerj.com (web)
--	--

application/pdf 2.4 MB
sha1:bca1531b0562c6d72e9c...

[web.archive.org \(webarchive\)](http://web.archive.org/webarchive/peerj.com)
[peerj.com \(web\)](http://peerj.com)

[Read Full Text](#)

Type article-journal
Stage published
Date 2018-02-13
DOI 10.7717/peerj.4375
PubMed 29456894
PMC PMC5815332
Wikidata Q49873702
Container Metadata
<ul style="list-style-type: none"> 🔖 Open Access Publication ✔ In DOAJ ✔ In ISSN ROAD 🔗 ISSN-L: 2167-8359 ➔ Fatcat Entry
Work Entity
grouping other versions (eg, pre-print) and variants of this release
▶ Lookup Links
Fatcat Bits
State is "active". Revision: de98deaf-4720-430a-8a97-1ff244cb602f
As JSON object via API
Edit Metadata View History

Fulltext
Mirror
Locations

Publication Status
PIDs Galore
Journal Metadata
Alt. Versions
API Helpers
Wiki Sauce



INTERNET ARCHIVE SCHOLAR

Search Millions of Research Papers

This fulltext search index includes over 25 million research articles and other scholarly documents preserved in the Internet Archive. The collection spans from digitized copies of eighteenth century journals though the latest Open Access conference proceedings and pre-prints crawled from the World Wide Web.

This service is in "alpha". It has several bugs, experiences downtime, and has not been officially announced.

<https://scholar.archive.org/>



blood clot

Search

User Guide

258,002 Hits in 0.91 sec

Release Date All Time Past Week Past Year Since 2000 Before 1925

Resource Type Papers Reports Datasets Everything Availability Fulltext Microfilm Open Access Metadata

Sort Order Relevancy Recent First Oldest First

Staghorn Blood Clot

Hsin-Ming Lee, Rheun-Chuan Lee, Wu-Chang Yang, Chih-Yu Yang 2011 *Internal medicine* (Tokyo. 1992)

Subsequent magnetic resonance imaging revealed the right renal pelvis and major calyces were filled with staghorn-shaped blood clots in both coronal sections of T1-weighted image (Picture A) and axial section of T2weighted image (Picture B). ... The term "staghorn blood clot" as first described by Ronchi et al (2), refers to a blood clot filling the collecting system and it results in obstructive uropathy. The authors state that they have no Conflict of Interest (COI). References 1. Lameire N, Hoste E. ... Subsequent magnetic resonance imaging revealed the right renal pelvis and major calyces were filled with staghorn-shaped blood clots in both coronal sections of T1-weighted image (Picture A) and axial section of T2weighted image (Picture B). ... The term "staghorn blood clot" as first described by Ronchi et al (2), refers to a blood clot filling the collecting system and it results in obstructive uropathy. The authors state that they have no Conflict of Interest (COI). ...

doi:10.2169/internalmedicine.50.6411 pmid:22041398 fatcat:ogibpbt2gjevfvdcpz2conqs33m



Blood platelets and blood clotting

T. F. Zucker 1913 *Experimental biology and medicine*

That the formed elements of the blood play a part in normal coagulation has long been known. Both leucocytes and platelets have been said to yield substances which contribute to fibrin formation. Leucocytes alone, however, will not coagulate fibrinogen. ... Cramer and Pringle' have recently shown that oxalate plasma freed from platelets by filtering through clay filters does not clot on adding an amount of CaCl2 which causes a similar centrifuged but unfiltered plasma to clot in a short time. ... Blood platelets and blood clotting. By T. F. ZUCKER (by invitation). [From the H. K. Cusling Laboratory of Exferimental Medicine, Western Reserve University, Cleveland, Ohio.] That the formed elements of the blood play a part in normal coagulation has long been known. ... Cramer and Pringle' have recently shown that oxalate plasma freed from platelets by filtering through clay filters does not clot on adding an amount of CaCl2 which causes a similar centrifuged but unfiltered plasma to clot in a short time. ...

doi:10.3181/00379727-11-36 fatcat:o34k6uydhbeim6dss5blsg7ly



web.archive.org



archive.org

<https://scholar.archive.org/>



ARCHIVE

The
DEMO
is here!!

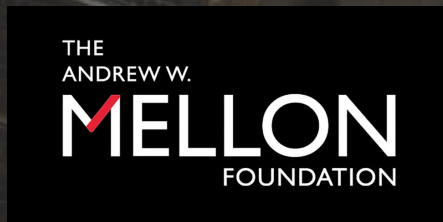
Check it out - 122MB

INTERNET



Current Work & Roadmap & Services

- **Beyond PDF: html, data, code, etc**
- **Secondaries & platforms**
- **“Save Paper Now”**
- **Full-text search for everything**
- **Citation integrity & indexing**
- **Bulk Data Sharing**



"Papers"	Total	118,822,564
	Fulltext on web	29,679,102
	"Gold" Open Access	20,176,845
	In a Keepers/KBART archive	73,528,229
	On web, not in Keepers	13,728,054
Releases	Total	166,085,403
	References (raw, unlinked)	1,107,532,083
Containers	Total	184,106



<https://fatcat.wiki/stats>

IIPC 2021

THANKS! CONTACT US!

Jefferson Bailey

Director, Web Archiving & Data Services,

jefferson@archive.org

Bryan Newbold, Open Data Engineer

bnewbold@archive.org

Special Thanks!!

Volunteers: David Rosenthal, Vicky Reich, Ellen Spertus

Funders: Mellon Foundation, IMLS



Internet Archive, <https://archive.org>

Fatcat beta, <https://fatcat.wiki/>

IIPC 2021