AUTOMATING PROCESSING AND INTAKE IN THE INSTITUTIONAL REPOSITORY WITH PYTHON

John Beatty

University at Buffalo
School of Law

University at Buffalo
School of Law

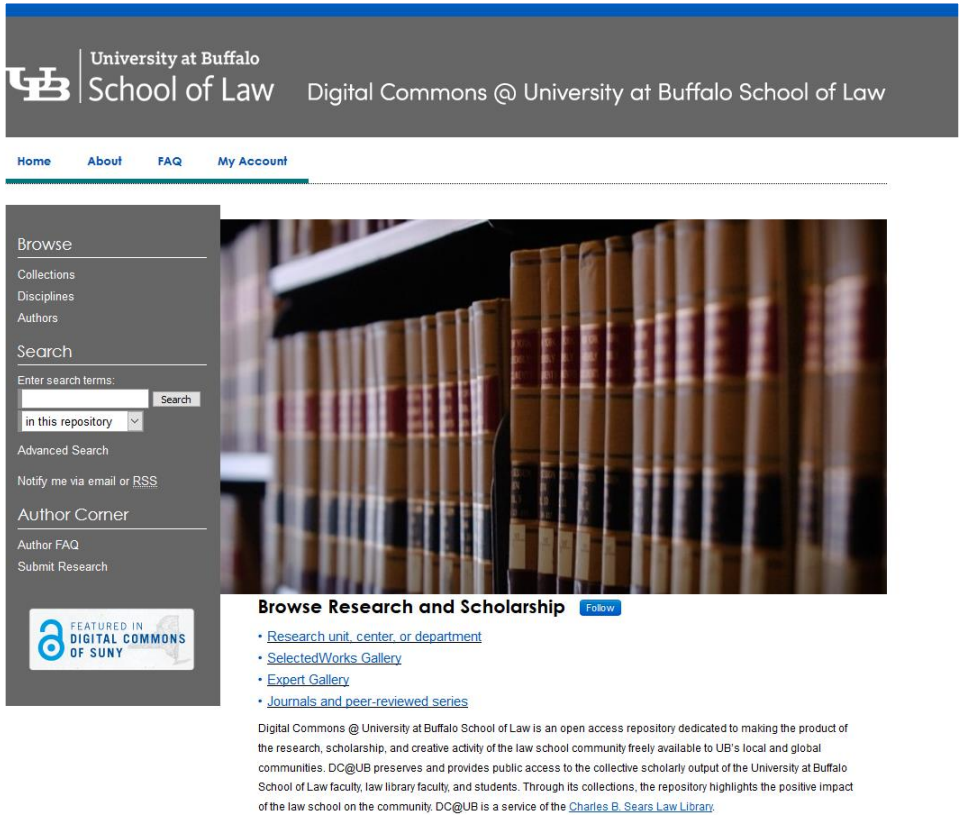
INTRODUCTION

What are we doing here?

University at Buffalo
School of Law

Populating the Institutional Repository

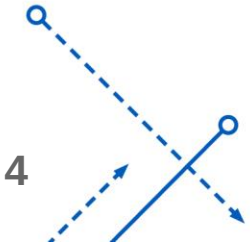
- Year 1: Faculty scholarship
- Year 2: Law Journals & Alumni Publications



Disclaimer/Full Disclosure

It's true I had zero Python programming knowledge at the start of this project. But I was starting with some knowledge:

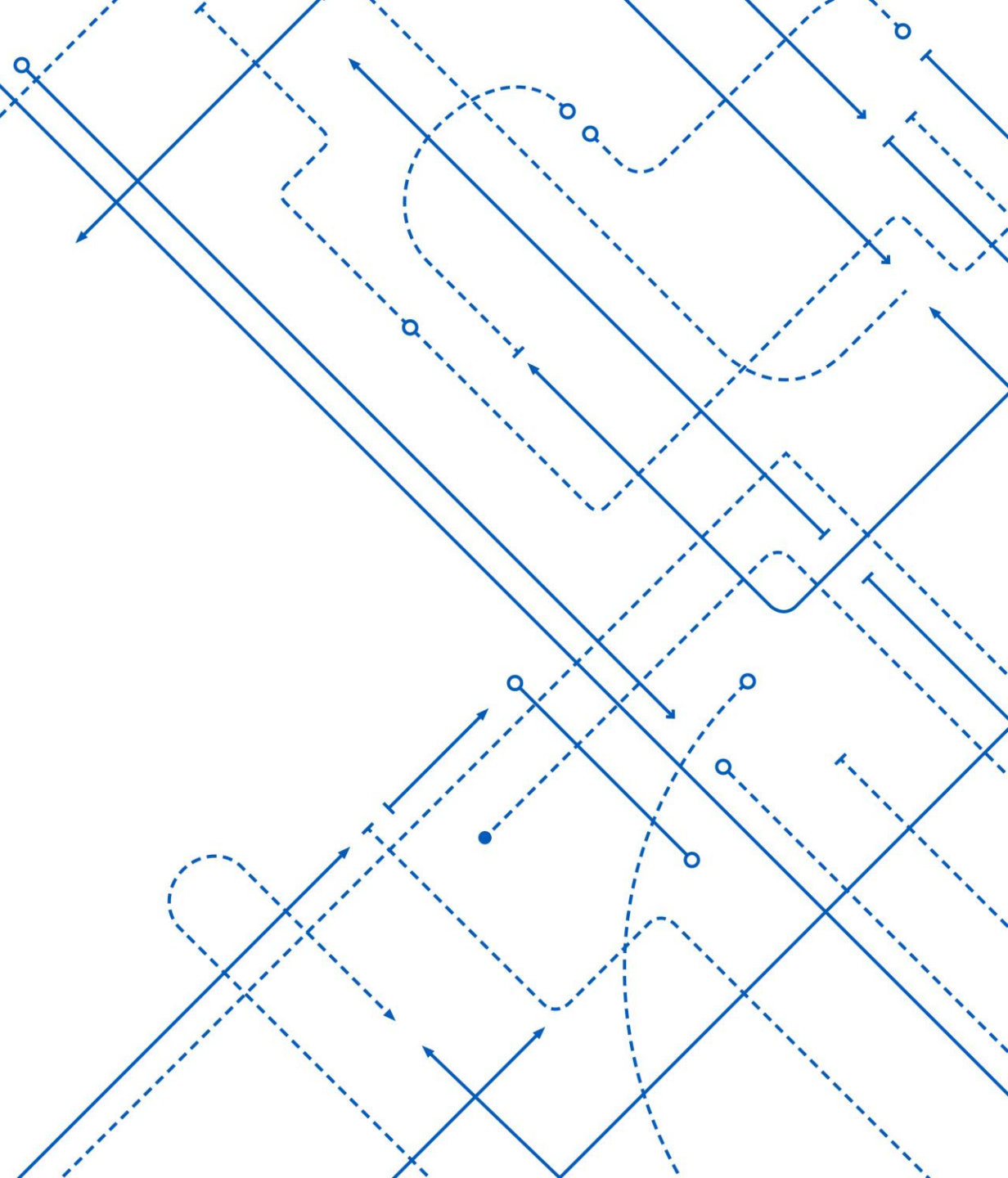
- General programming knowledge & experience
- VisualBasic (17 years; 4 database applications built)
- Bash (basic knowledge)
- Perl (mostly forgotten knowledge)
- Has used a regular expression



University at Buffalo
School of Law

JOURNALS

The Law Journal Project

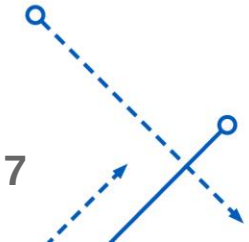


The Law Journal Project: The Journals

- Buffalo Environmental Law Review: 23 volumes, 1-2 issues/volume
- Buffalo Human Rights Law Review: 22 volumes, 1-2 issues/volume
- Buffalo Intellectual Property Journal: 11 volumes, 1-2 issues/volume
- Buffalo Journal of Gender, Law & Social Policy: 24 volumes, 1 issue/volume
- Buffalo Law Review: 65 volumes, 3-5 issues/volume
- Buffalo Public Interest Law Review: 35 volumes, 1 issue/volume

The Law Journal Project: Workflow

- Convert Hein metadata to Digital Commons format
- Load PDFs into Box drive
- Preview files in Box
- Check metadata against PDF and correct where necessary
- Cut and paste Box links into Digital Commons spreadsheet
- Upload



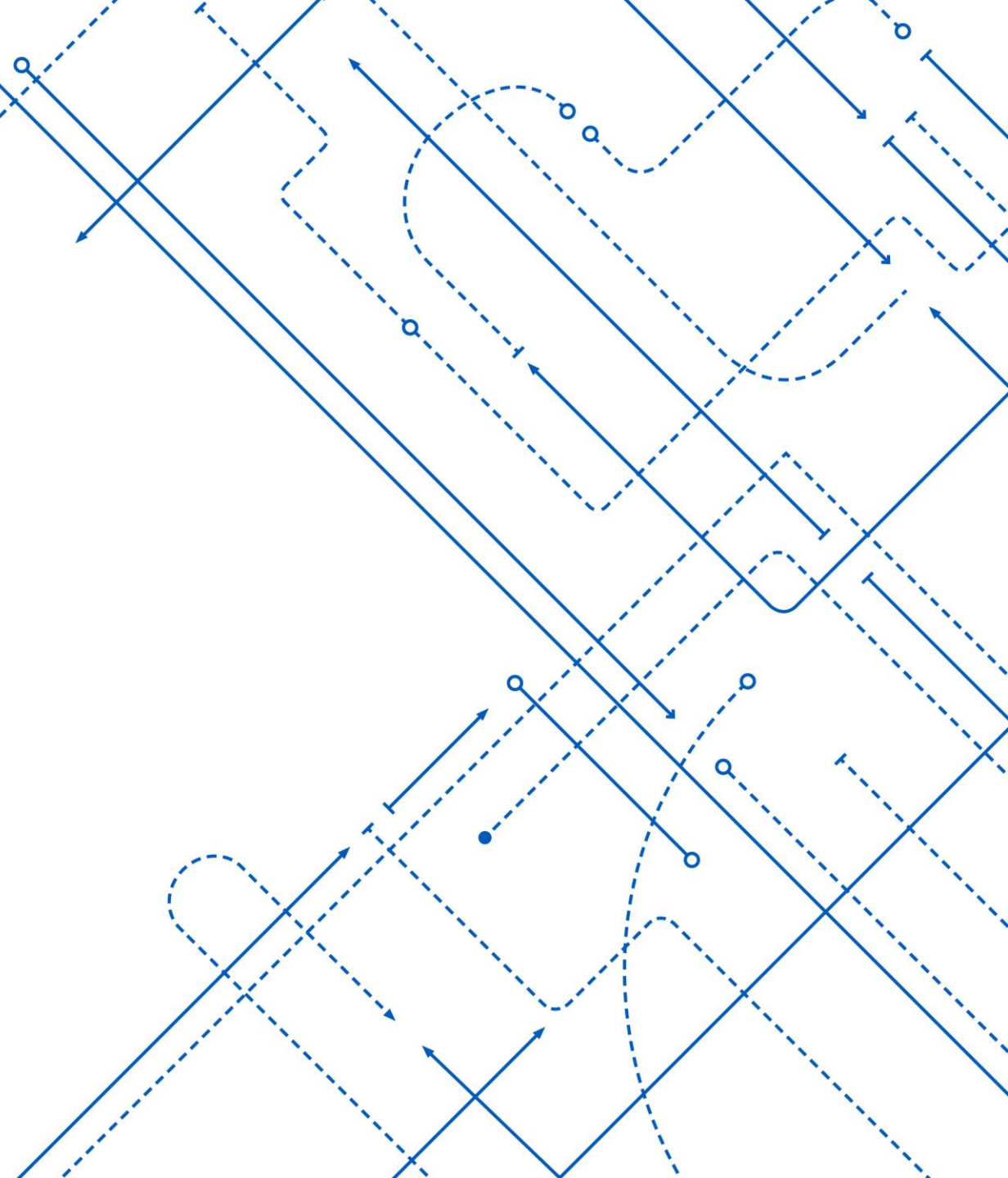
The Law Journal Project: Timeline

- August-November 2018
- All but first 22 volumes of Buffalo Law Review complete in mid-November

University at Buffalo
School of Law

THE PROBLEM

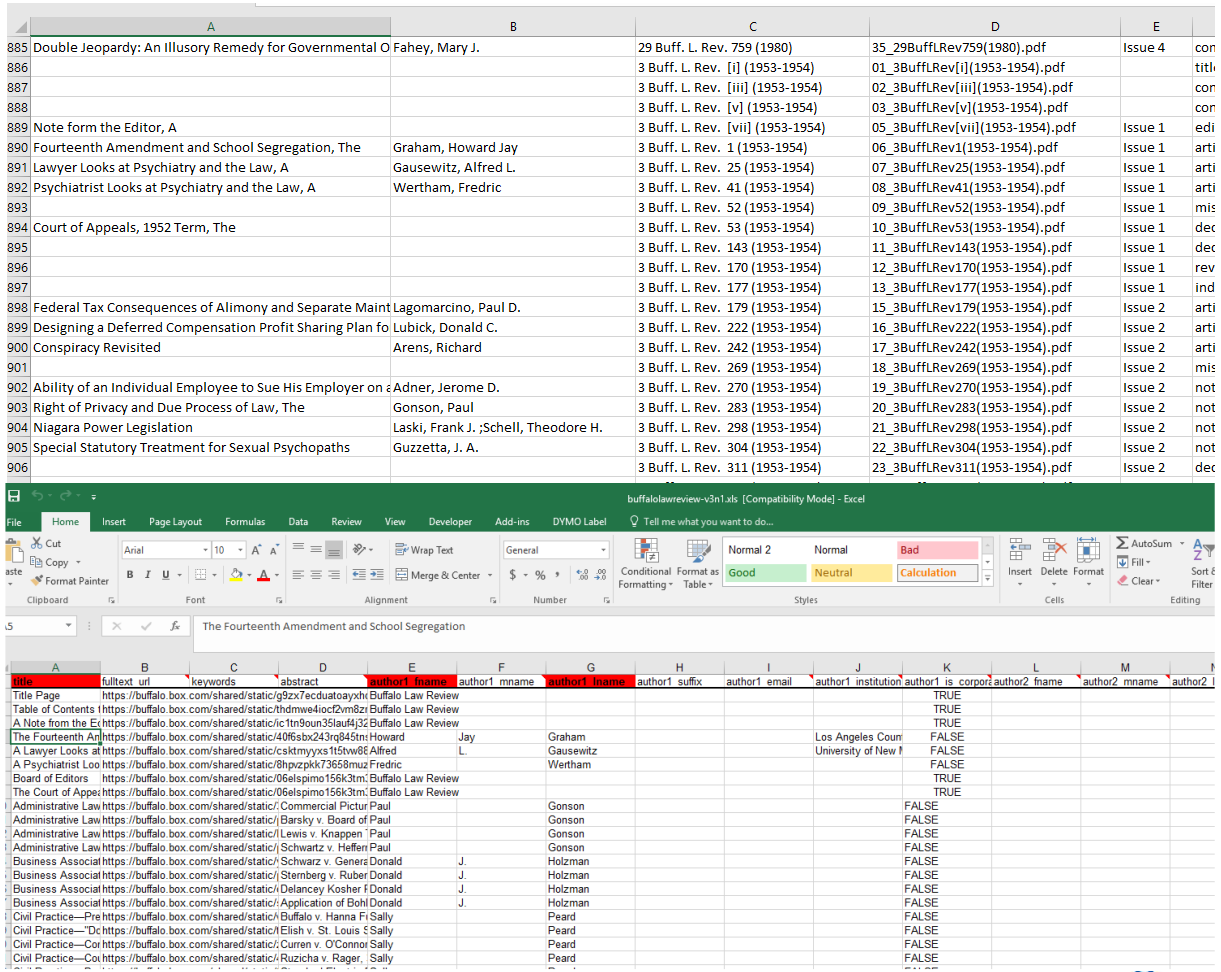
What's so special about the Law Review?



University at Buffalo
School of Law

Conversion from HeinOnline to IR

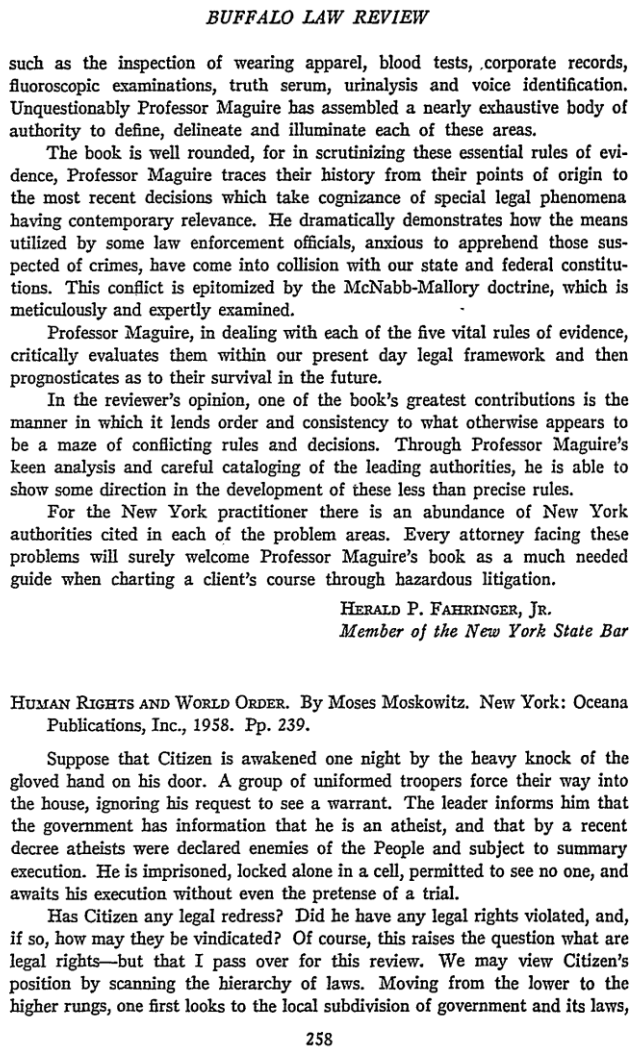
- Some types of documents are in the system as a section rather than individual pieces
- Combined files have no individual metadata
- Some documents have no author data
- Some articles missing the last page



	A	B	C	D	E	
885	Double Jeopardy: An Illusory Remedy for Governmental O	Fahey, Mary J.	29 Buff. L. Rev. 759 (1980)	35_29BuffLRev759(1980).pdf	Issue 4	con
886			3 Buff. L. Rev. [i] (1953-1954)	01_3BuffLRev[i](1953-1954).pdf		titl
887			3 Buff. L. Rev. [iii] (1953-1954)	02_3BuffLRev[iii](1953-1954).pdf		con
888			3 Buff. L. Rev. [v] (1953-1954)	03_3BuffLRev[v](1953-1954).pdf		con
889	Note form the Editor, A		3 Buff. L. Rev. [vii] (1953-1954)	05_3BuffLRev[vii](1953-1954).pdf	Issue 1	edi
890	Fourteenth Amendment and School Segregation, The	Graham, Howard Jay	3 Buff. L. Rev. 1 (1953-1954)	06_3BuffLRev1(1953-1954).pdf	Issue 1	arti
891	Lawyer Looks at Psychiatry and the Law, A	Gausewitz, Alfred L.	3 Buff. L. Rev. 25 (1953-1954)	07_3BuffLRev25(1953-1954).pdf	Issue 1	arti
892	Psychiatrist Looks at Psychiatry and the Law, A	Wertham, Fredric	3 Buff. L. Rev. 41 (1953-1954)	08_3BuffLRev41(1953-1954).pdf	Issue 1	arti
893			3 Buff. L. Rev. 52 (1953-1954)	09_3BuffLRev52(1953-1954).pdf	Issue 1	mis
894	Court of Appeals, 1952 Term, The		3 Buff. L. Rev. 53 (1953-1954)	10_3BuffLRev53(1953-1954).pdf	Issue 1	dec
895			3 Buff. L. Rev. 143 (1953-1954)	11_3BuffLRev143(1953-1954).pdf	Issue 1	dec
896			3 Buff. L. Rev. 170 (1953-1954)	12_3BuffLRev170(1953-1954).pdf	Issue 1	rev
897			3 Buff. L. Rev. 177 (1953-1954)	13_3BuffLRev177(1953-1954).pdf	Issue 1	ind
898	Federal Tax Consequences of Alimony and Separate Maint	Lagomarcino, Paul D.	3 Buff. L. Rev. 179 (1953-1954)	15_3BuffLRev179(1953-1954).pdf	Issue 2	arti
899	Designing a Deferred Compensation Profit Sharing Plan fo	Lubick, Donald C.	3 Buff. L. Rev. 222 (1953-1954)	16_3BuffLRev222(1953-1954).pdf	Issue 2	arti
900	Conspiracy Revisited	Arens, Richard	3 Buff. L. Rev. 242 (1953-1954)	17_3BuffLRev242(1953-1954).pdf	Issue 2	arti
901			3 Buff. L. Rev. 269 (1953-1954)	18_3BuffLRev269(1953-1954).pdf	Issue 2	mis
902	Ability of an Individual Employee to Sue His Employer on a	Adner, Jerome D.	3 Buff. L. Rev. 270 (1953-1954)	19_3BuffLRev270(1953-1954).pdf	Issue 2	not
903	Right of Privacy and Due Process of Law, The	Gonson, Paul	3 Buff. L. Rev. 283 (1953-1954)	20_3BuffLRev283(1953-1954).pdf	Issue 2	not
904	Niagara Power Legislation	Laski, Frank J. ;Schell, Theodore H.	3 Buff. L. Rev. 298 (1953-1954)	21_3BuffLRev298(1953-1954).pdf	Issue 2	not
905	Special Statutory Treatment for Sexual Psychopaths	Guzzetta, J. A.	3 Buff. L. Rev. 304 (1953-1954)	22_3BuffLRev304(1953-1954).pdf	Issue 2	not
906			3 Buff. L. Rev. 311 (1953-1954)	23_3BuffLRev311(1953-1954).pdf	Issue 2	dec

Book Reviews

- In HeinOnline, book reviews in a single BLR issue are all in one file
- All book reviews are signed, but no author data in HeinOnline
- In later volumes (processed first), issues contain 2 book reviews at most; splitting and metadata creation was done by hand
- In early volumes, there are up to five book reviews an issue, so automation helpful



such as the inspection of wearing apparel, blood tests, corporate records, fluoroscopic examinations, truth serum, urinalysis and voice identification. Unquestionably Professor Maguire has assembled a nearly exhaustive body of authority to define, delineate and illuminate each of these areas.

The book is well rounded, for in scrutinizing these essential rules of evidence, Professor Maguire traces their history from their points of origin to the most recent decisions which take cognizance of special legal phenomena having contemporary relevance. He dramatically demonstrates how the means utilized by some law enforcement officials, anxious to apprehend those suspected of crimes, have come into collision with our state and federal constitutions. This conflict is epitomized by the McNabb-Mallory doctrine, which is meticulously and expertly examined.

Professor Maguire, in dealing with each of the five vital rules of evidence, critically evaluates them within our present day legal framework and then prognosticates as to their survival in the future.

In the reviewer's opinion, one of the book's greatest contributions is the manner in which it lends order and consistency to what otherwise appears to be a maze of conflicting rules and decisions. Through Professor Maguire's keen analysis and careful cataloging of the leading authorities, he is able to show some direction in the development of these less than precise rules.

For the New York practitioner there is an abundance of New York authorities cited in each of the problem areas. Every attorney facing these problems will surely welcome Professor Maguire's book as a much needed guide when charting a client's course through hazardous litigation.

HERALD P. FAHRINGER, JR.
Member of the New York State Bar

HUMAN RIGHTS AND WORLD ORDER. By Moses Moskowitz. New York: Oceana Publications, Inc., 1958. Pp. 239.

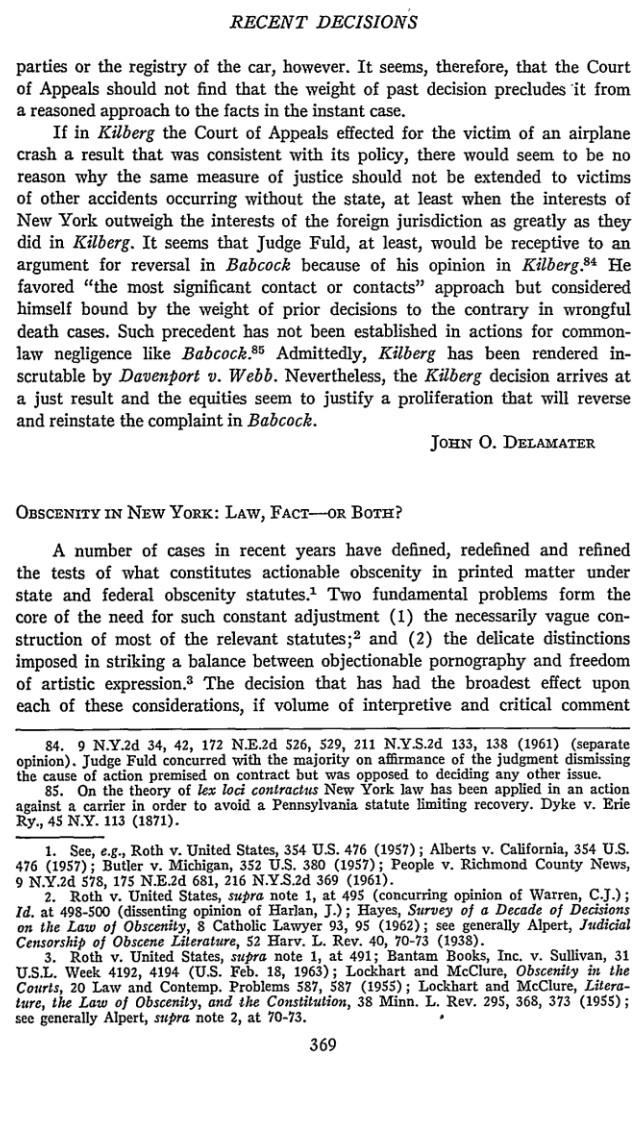
Suppose that Citizen is awakened one night by the heavy knock of the gloved hand on his door. A group of uniformed troopers force their way into the house, ignoring his request to see a warrant. The leader informs him that the government has information that he is an atheist, and that by a recent decree atheists were declared enemies of the People and subject to summary execution. He is imprisoned, locked alone in a cell, permitted to see no one, and awaits his execution without even the pretense of a trial.

Has Citizen any legal redress? Did he have any legal rights violated, and, if so, how may they be vindicated? Of course, this raises the question what are legal rights—but that I pass over for this review. We may view Citizen's position by scanning the hierarchy of laws. Moving from the lower to the higher rungs, one first looks to the local subdivision of government and its laws,

258

Case Notes & Legislative Notes

- Case notes all combined into all case notes for an entire issue
- No individual note or author metadata
- 2-3 issues/volume
- 5-10 case notes/issue
- Same for legislative notes, but only a few issues have them



RECENT DECISIONS

parties or the registry of the car, however. It seems, therefore, that the Court of Appeals should not find that the weight of past decision precludes it from a reasoned approach to the facts in the instant case.

If in *Kilberg* the Court of Appeals effected for the victim of an airplane crash a result that was consistent with its policy, there would seem to be no reason why the same measure of justice should not be extended to victims of other accidents occurring without the state, at least when the interests of New York outweigh the interests of the foreign jurisdiction as greatly as they did in *Kilberg*. It seems that Judge Fuld, at least, would be receptive to an argument for reversal in *Babcock* because of his opinion in *Kilberg*.[84] He favored "the most significant contact or contacts" approach but considered himself bound by the weight of prior decisions to the contrary in wrongful death cases. Such precedent has not been established in actions for common-law negligence like *Babcock*.[85] Admittedly, *Kilberg* has been rendered inscrutable by *Davenport v. Webb*. Nevertheless, the *Kilberg* decision arrives at a just result and the equities seem to justify a proliferation that will reverse and reinstate the complaint in *Babcock*.

JOHN O. DELAMATER

OBSCENITY IN NEW YORK: LAW, FACT—OR BOTH?

A number of cases in recent years have defined, redefined and refined the tests of what constitutes actionable obscenity in printed matter under state and federal obscenity statutes.[1] Two fundamental problems form the core of the need for such constant adjustment (1) the necessarily vague construction of most of the relevant statutes;[2] and (2) the delicate distinctions imposed in striking a balance between objectionable pornography and freedom of artistic expression.[3] The decision that has had the broadest effect upon each of these considerations, if volume of interpretive and critical comment

84. 9 N.Y.2d 34, 42, 172 N.E.2d 526, 529, 211 N.Y.S.2d 133, 138 (1961) (separate opinion). Judge Fuld concurred with the majority on affirmance of the judgment dismissing the cause of action premised on contract but was opposed to deciding any other issue.

85. On the theory of *lex loci contractus* New York law has been applied in an action against a carrier in order to avoid a Pennsylvania statute limiting recovery. Dyke v. Erie Ry., 45 N.Y. 113 (1871).

1. See, *e.g.*, Roth v. United States, 354 U.S. 476 (1957); Alberts v. California, 354 U.S. 476 (1957); Butler v. Michigan, 352 U.S. 380 (1957); People v. Richmond County News, 9 N.Y.2d 578, 175 N.E.2d 681, 216 N.Y.S.2d 369 (1961).

2. Roth v. United States, *supra* note 1, at 495 (concurring opinion of Warren, C.J.); *Id.* at 498-500 (dissenting opinion of Harlan, J.); Hayes, *Survey of a Decade of Decisions on the Law of Obscenity*, 8 Catholic Lawyer 93, 95 (1962); see generally Alpert, *Judicial Censorship of Obscene Literature*, 52 Harv. L. Rev. 40, 70-73 (1938).

3. Roth v. United States, *supra* note 1, at 491; Bantam Books, Inc. v. Sullivan, 31 U.S.L. Week 4192, 4194 (U.S. Feb. 18, 1963); Lockhart and McClure, *Obscenity in the Courts*, 20 Law and Contemp. Problems 587, 587 (1955); Lockhart and McClure, *Literature, the Law of Obscenity, and the Constitution*, 38 Minn. L. Rev. 295, 368, 373 (1955); see generally Alpert, *supra* note 2, at 70-73.

369

Court of Appeals

- Court of Appeals is highest court in New York
- Volumes 3-14 contain case note summaries for the prior year's Court of Appeals term
- 1 or 2 issues/volume
- Up to 150 case notes/issue
- In most volumes, notes are signed

COURT OF APPEALS, 1956 TERM

ADMINISTRATIVE LAW

Board Member Absent From Administrative Hearing Entitled To Vote

Courts, in reviewing the decisions of administrative tribunals, are limited to ascertaining whether or not there is a rational, legal basis for such determinations.[1] This does not mean that the courts will "Probe the mental processes" of the body[2] but rather that they will determine only whether there has been an independent appraisal and conclusion based on sufficient knowledge.[3]

In *Taub v. Pirnie*,[4] an application for a variance from a local zoning ordinance, the Court of Appeals unanimously held that, though a member of the Board was absent from the public hearing and did not read the transcript of the record, he nevertheless possessed adequate knowledge to make an informed decision. The Court pointed out that because the absent member had for years been a resident of the village and a village trustee, and because he had discussed the arguments presented at the hearing before voting, he possessed sufficient information to reach an independent and competent conclusion.

Thus the Court has held that an administrative board member may vote on the disposition of a proceeding, though absent from the hearing thereon, provided that he has acquired independent knowledge of the issues involved. It is submitted that this proposition is susceptible of unwarranted extension and its application should be restricted to similar factual settings.

Hotels Not Special Class Of Housing Accomodations Within Rent Control Law

In *Hotel Association of New York City v. Weaver*[5] the Court held that the Temporary State Housing Commission did not act in an arbitrary and capricious manner when it ruled that hotels as such did not constitute a "particular class of housing accommodations" within the meaning of the Emergency Housing Rent Control Law.[6]

Petitioner, an association of one hundred and seventy-one hotels in New York City, petitioned the Rent Administrator to find that hotels as such are a particular class of housing accommodations within the meaning of the statute and entitled

1. Mounting & Finishing Co. vs McGoldrick, 294 N.Y. 104, 60 N.E.2d 825 (1945).
2. Kilgus v. Board of Estimate of the City of New York, 308 N.Y. 620,628, 127 N.E.2d 705, 710 (1955).
3. *Ibid.*
4. 3 N.Y.2d 188, 165 N.Y.S.2d 1 (1957).
5. 3 N.Y.2d 206, 165 N.Y.S.2d 17 (1957).
6. N.Y. EMERGENCY HOUSING RENT CONTROL LAW §12.

63

Student Notes & Comments

- In early volumes, notes did not always start at the top of a page
- All page breaks were at the start of the next note
- Some notes missing the last page

Supreme Court taking jurisdiction. This would eliminate the risk of juries running wild and at the same time determine fairly the merits of a citizen's claim. As it now stands liability of the city for all practical purposes is *dead*, and the individual claimant is being denied a common law right to bring successful suit against a municipal corporation for its tortious neglect to maintain the streets and sidewalks in a reasonable safe condition for travel.

Anthony J. Colucci, Jr.

VALUATION OF PUBLIC UTILITY PROPERTY FOR TAX PURPOSES: THE NEW YORK SPECIAL FRANCHISE TAX

"What is a cynic? A man who knows the price of everything and the value of nothing."

—Oscar Wilde.

Whether the jaundiced outlook of a habitual cynic is the cause producing the regrettable imbalance noted above, or whether a prolonged struggle to discover a rational relationship between the opposed concepts might be the cause of cynicism, the fact remains that when the ineluctable exigencies of taxation require the expression of *value* in dollars and cents, which are properly the terms of *price*, problems are generated the subtlety of which may well threaten the most equable temperament.

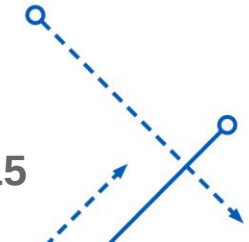
Public utilities share with all other commercial and industrial enterprises many general problems of this character, and such problems, since they are not peculiar to public utilities, are beyond the scope of this paper.

The public utility encounters its own special brand of problem in two areas affected by property valuation: real property taxation and rate-making. This paper will concern itself mainly with the first of these areas, although some adversions to the other will be necessary with respect to the question of the evidentiary merit, in proceedings arising out of contested assessments, of valuations made by the Public Service Commission for rate-making purposes.

One of the distinguishing features of a public utility producing valuation problems peculiar to an enterprise of this kind is its statutory right to use public streets and lands in ways which, without such authority, would constitute trespass. Gas, electric, and water companies are good examples, since they clearly must run pipes and wires through streets and highways in order to serve their

Why do the extra work?

- In some cases, combined works are substantial (review essays)
- To properly credit alumni and faculty authors
- Some case notes are contemporary coverage of substantial changes in New York or United States law
- Some notes written by prominent alumni



Implementation Issues

Previous Solution

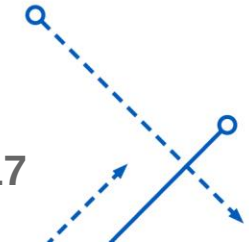
- Requires 2 librarians and a student worker

Our Situation

- Tech services departments busy with massive LSP migration
- Most departments shorthanded because of retirement
- No funding for student workers

The Solution: Automation

- Personnel available: 1 Faculty Scholarship Librarian
- Drastically shorten the amount of time needed to generate metadata and split PDFs
- Use generated metadata and split PDFs in established workflow



Timelines

Proposed:

- Learn enough Python to start coding: 1-2 weeks
- Write initial code and test: 1-2 weeks
- Process 22 volumes: 1 month

Actual:

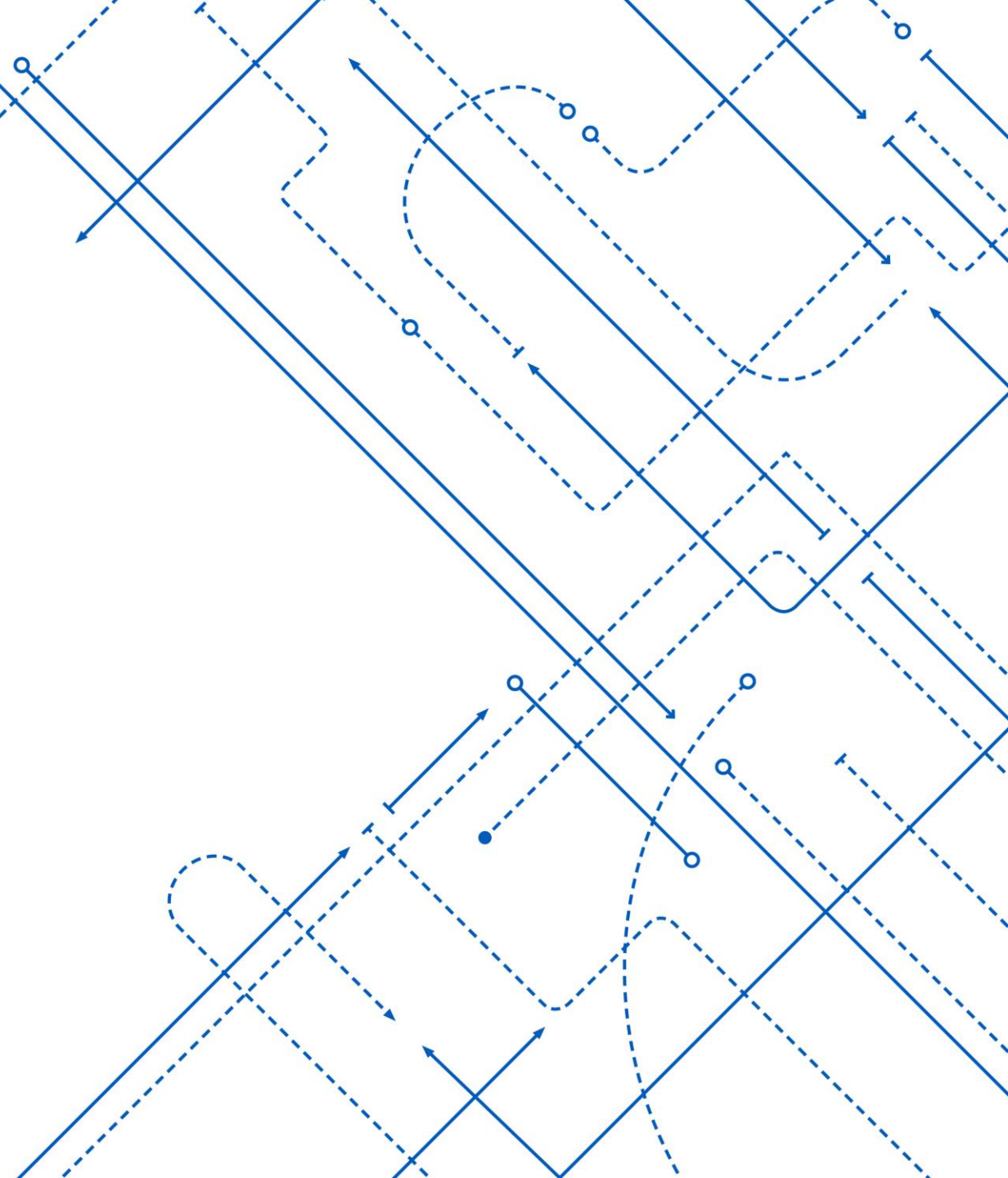
- Learn enough Python to start: 3 days (Thanksgiving week)
- Initial code and test: 5 days (November 26-30)
- Process 22 volumes: 4 weeks (December 3-21, January 3-11)

Note: Processing time included a **LOT** of code tweaking.

University at Buffalo
School of Law

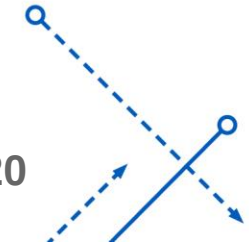
THE PROJECT

First Steps



Learning Python

- John Mueller: Beginning Programming with Python for Dummies
- Kent D. Lee: Python Programming Fundamentals
- T.R. Padmanabhan: Programming with Python
- Python Documentation: https://docs.python.org/3/
- w3schools.com: https://www.w3schools.com/python/default.asp
- Automate the Boring Stuff: https://automatetheboringstuff.com/



Programming Environment

- Laptop computer running Ubuntu Linux 18.04
- PyCharm Community Edition (free!)
- Python 3.6

Identifying Libraries

- PyPDF2: PDF toolkit that can be used to extract data and manipulate PDF files
- pdfminer: A tool for extracting information from PDF files (using pdfminer.six, for Python 3 compatibility)
- openpyxl: Python library to read and write Excel 2010 xlsx/xlsm files
- Standard Python libraries: argparse, os, re, csv, fnmatch, io
- Add-on libraries installed with PIP

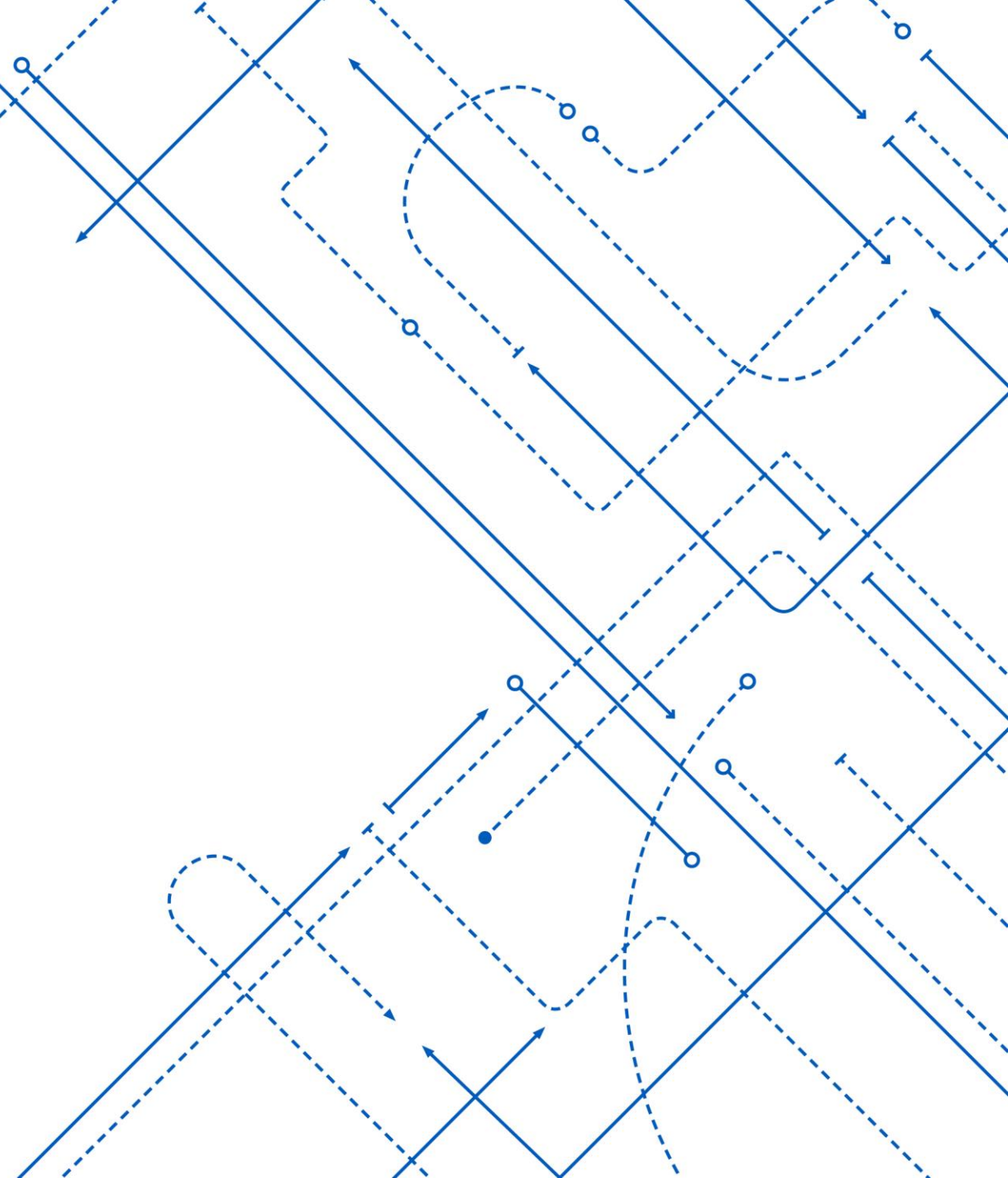
Wait… TWO PDF libraries?

- Yes, two PDF libraries
- PyPDF2 has good tools for manipulating PDFs, but the documentation specifically says not to rely on the text extraction functions
- pdfminer is designed to extract information including text and layout from PDF files, so can be relied on for text extraction. But it doesn't have the manipulation functions.

University at Buffalo
School of Law

THE PROJECT

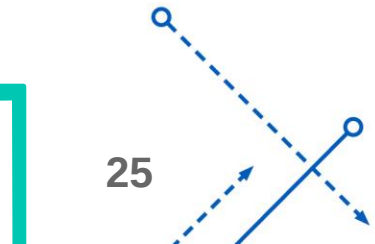
Workflow



Initial Workflow: Single Script

- Search through PDF for start page (PDF), end page (PDF), author, title, start page (printed)
- Split PDF into multiple files based on start and end pages
- Export metadata into Excel file to be cut and pasted into Digital Commons batch spreadsheets

Just One Problem:
OCR. It's not good enough to allow the code to consistently identify the metadata elements.



Scan file

Use appropriate dsplit-XX.py to extract metadata. Use the --write-csv-only option because none of the OCR is good enough to trust that it's right.

```
eview/buflr1/09_1BuffLRev28(1951-1952).pdf
john@john-UX31A:~/PycharmProjects/journaltools$ python3 dsplit-rd.py -v --write-
csv-only '/media/john/B48A-17E4/Buffalo Law Review/buflr1/09_1BuffLRev28(1951-19
52)-NEW.pdf'
Reading file: /media/john/B48A-17E4/Buffalo Law Review/buflr1/09_1BuffLRev28(195
1-1952)-NEW.pdf

Processing PDF pages
WARNING! Missing titles, start pages, and starting PDF pages
Data written to file: /media/john/B48A-17E4/Buffalo Law Review/buflr1/09_1BuffLR
ev28(1951-1952)-NEW.csv
john@john-UX31A:~/PycharmProjects/journaltools$
```

09_1BuffLRev28(1951-1952)-NEW.csv - LibreOffice Calc

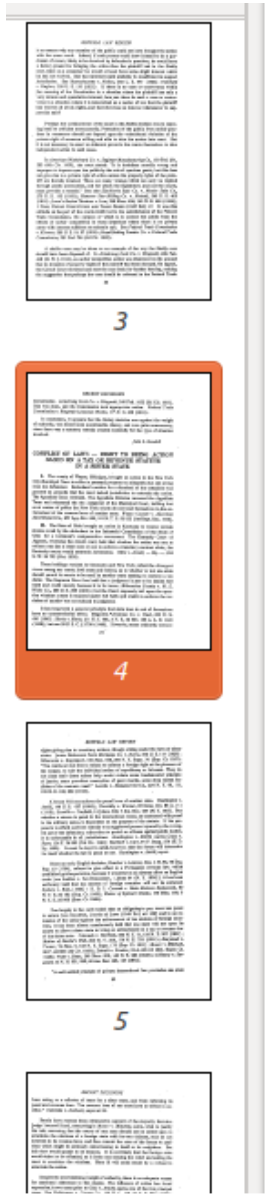
File Edit View Insert Format Styles Sheet Data Tools Window Help

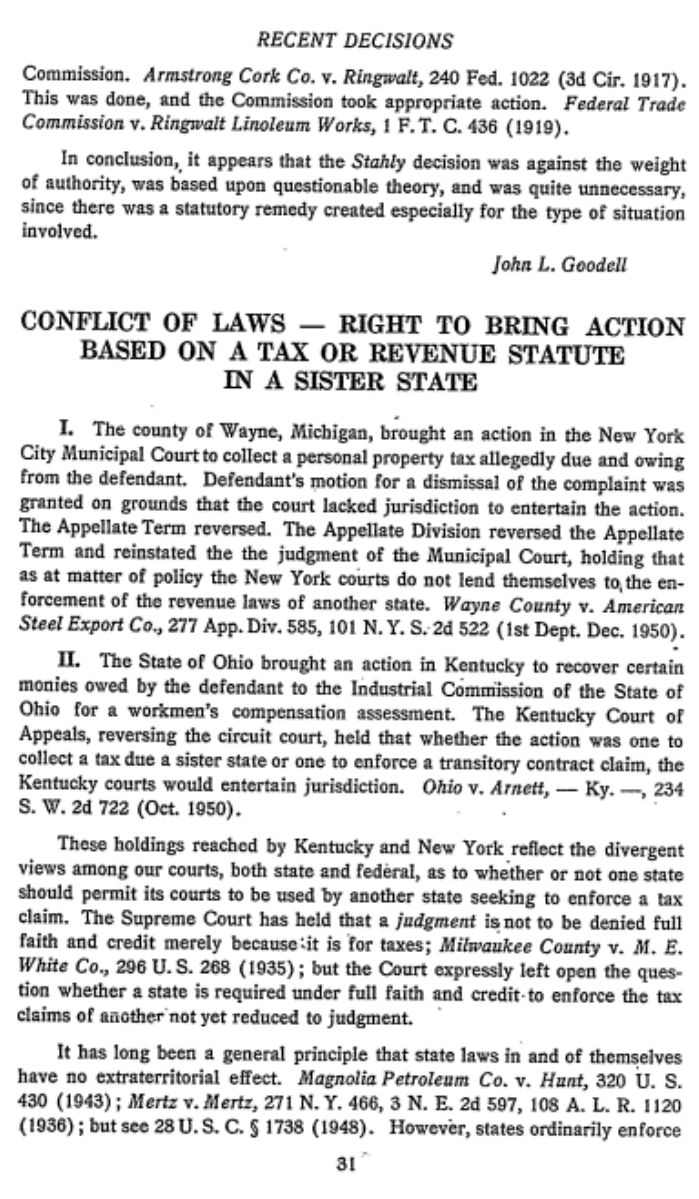
A1 | title

	A	B	C	D	E	F	G	H	I
1	title	start_page	start_pdf_page	end_pdf_page	f_name1	m_name1	l_name1	suffix1	f_name2
2	Unfair Competition public Will Be Protected Although Plaintiff Has No Standing in Court	28	0	3	John	L.	Goodell		In
3	Conflict of Laws right to Bring Action Based on A Tax or Revenue Statute in A Sister State	31	3	6	Alvin	M.	Glick		
4	Labor Law injunctions jurisdiction of N. L. R. B. Held Exclusive		7	11	Francis	W.	Greune,		
5	Wills totten Trusts distribution of Balance of Deposit After Invasion by Widow		12	15	Samuel	R.	Miserendino		Robert
6	Administrative Lawworkmen's Compensation "Accidental Injury" as A Question Of. Fact	43	15	18	Robert	B.	Fleming		
7	Taxation federal Income Tax interest in Law Partnership as A Capital Asset		18	21	Henry		Woicicki		
8	Conflict of Laws-Devise of Foreign Realty renvoi Expressly Adopted	49	21	24	Philip	A.	Erickson		
9	Husband and Wife recovery by Wife for Loss of Consortium	52	24	26	Thomas	J.	Kelly		
10	Annulment Proceedings for Fraud sufficiency of Proof how to Prove Refusal to Have Sexual Intercourse		26	28	Morree	M.	Levine		
11	Negligence last Clear Chance Held Applicable Though Defendant's Driver Did Not See Decedent		28	31	Alvin	M.	Glick		
12	Limitation of Actions implied Warranty "Contract" Action Held Barred by Injury to Property Statute of Limitations		31	33	John	G.	Wick		
13	Labor Law discrimination by Union Seeking Union Shop on Basis of Sex Held "Unreasonable"	61	33	34	In	Wilson	V. Hacker,		
14	Municipal Corporationsnotice of Claim infants		36	36	David		Buch		
15	Unfair Competition fair Trade Law Held to Prohibit Cash Register Receipt Discount Plan		38	38	Burton	B.	Sarles		
16	Constitutional Law-Civil Rights - "Threat" to Public Order Held Superior to Freedom of Speech	71	40	40	Mary	K.	Davey		
17	Liability of Remote Supplier		42	42	Edward	S.	Spector		
18	Workmen's Compensation injuries Arising out of and in Course of Employment when Does Horseplay Amount to Abandonment	74	46	46	Phyllis	J.	Hubbard		David
19			0	49	Philip	A.	Erickson		
20									
21									

Check metadata

Open the CSV file and check it against the original PDF. Fix titles, authors, and most importantly, start and ending pages for the PDF split.





RECENT DECISIONS

Commission. *Armstrong Cork Co.* v. *Ringwalt*, 240 Fed. 1022 (3d Cir. 1917). This was done, and the Commission took appropriate action. *Federal Trade Commission* v. *Ringwalt Linoleum Works*, 1 F. T. C. 436 (1919).

In conclusion, it appears that the *Stahly* decision was against the weight of authority, was based upon questionable theory, and was quite unnecessary, since there was a statutory remedy created especially for the type of situation involved.

John L. Goodell

CONFLICT OF LAWS — RIGHT TO BRING ACTION BASED ON A TAX OR REVENUE STATUTE IN A SISTER STATE

I. The county of Wayne, Michigan, brought an action in the New York City Municipal Court to collect a personal property tax allegedly due and owing from the defendant. Defendant's motion for a dismissal of the complaint was granted on grounds that the court lacked jurisdiction to entertain the action. The Appellate Term reversed. The Appellate Division reversed the Appellate Term and reinstated the the judgment of the Municipal Court, holding that as at matter of policy the New York courts do not lend themselves to the enforcement of the revenue laws of another state. *Wayne County* v. *American Steel Export Co.*, 277 App. Div. 585, 101 N. Y. S. 2d 522 (1st Dept. Dec. 1950).

II. The State of Ohio brought an action in Kentucky to recover certain monies owed by the defendant to the Industrial Commission of the State of Ohio for a workmen's compensation assessment. The Kentucky Court of Appeals, reversing the circuit court, held that whether the action was one to collect a tax due a sister state or one to enforce a transitory contract claim, the Kentucky courts would entertain jurisdiction. *Ohio* v. *Arnett*, — Ky. —, 234 S. W. 2d 722 (Oct. 1950).

These holdings reached by Kentucky and New York reflect the divergent views among our courts, both state and federal, as to whether or not one state should permit its courts to be used by another state seeking to enforce a tax claim. The Supreme Court has held that a *judgment* is not to be denied full faith and credit merely because it is for taxes; *Milwaukee County* v. *M. E. White Co.*, 296 U. S. 268 (1935); but the Court expressly left open the question whether a state is required under full faith and credit to enforce the tax claims of another not yet reduced to judgment.

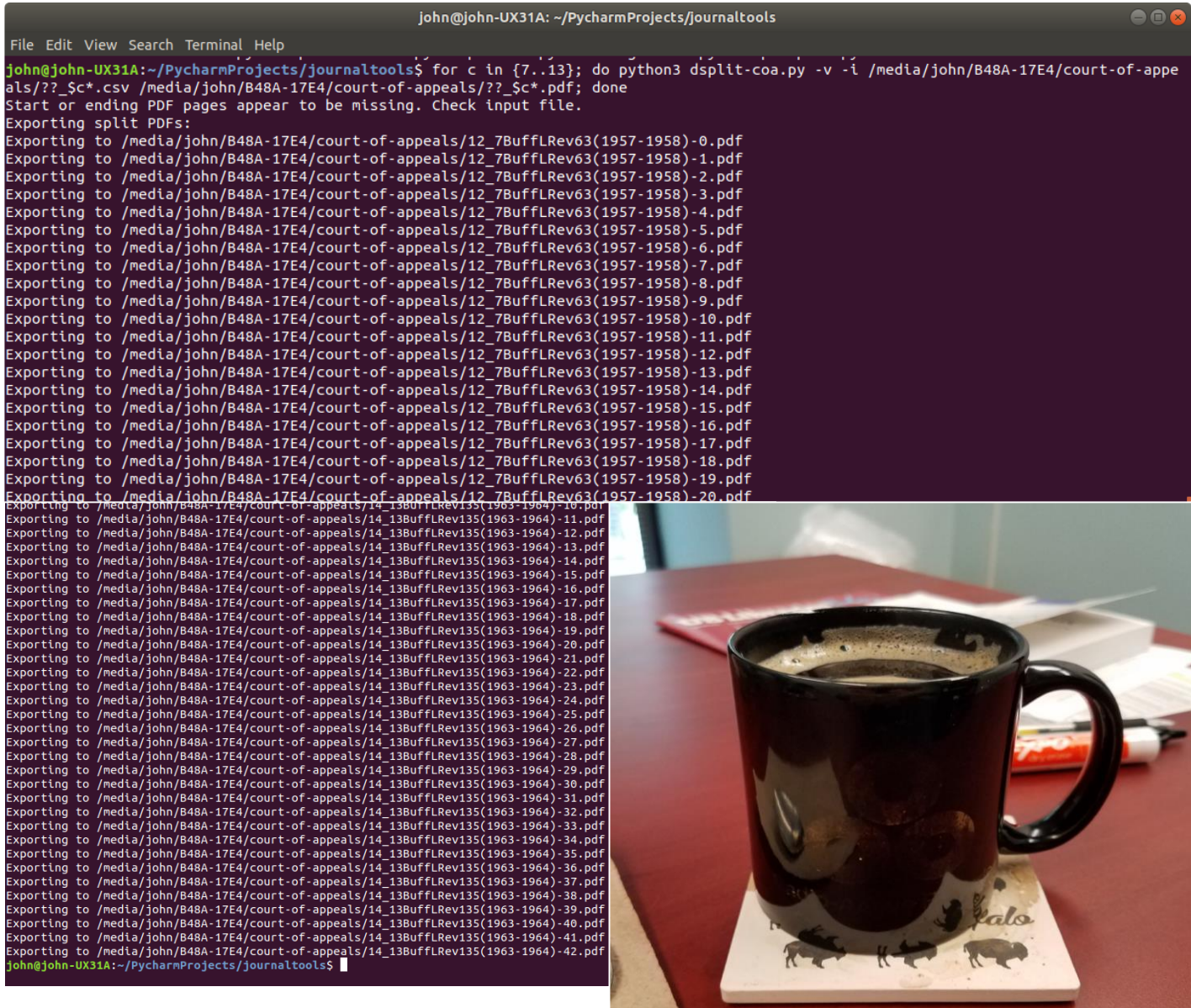
It has long been a general principle that state laws in and of themselves have no extraterritorial effect. *Magnolia Petroleum Co.* v. *Hunt*, 320 U. S. 430 (1943); *Mertz* v. *Mertz*, 271 N. Y. 466, 3 N. E. 2d 597, 108 A. L. R. 1120 (1936); but see 28 U. S. C. § 1738 (1948). However, states ordinarily enforce

31

University at Buffalo
School of Law

Split PDF

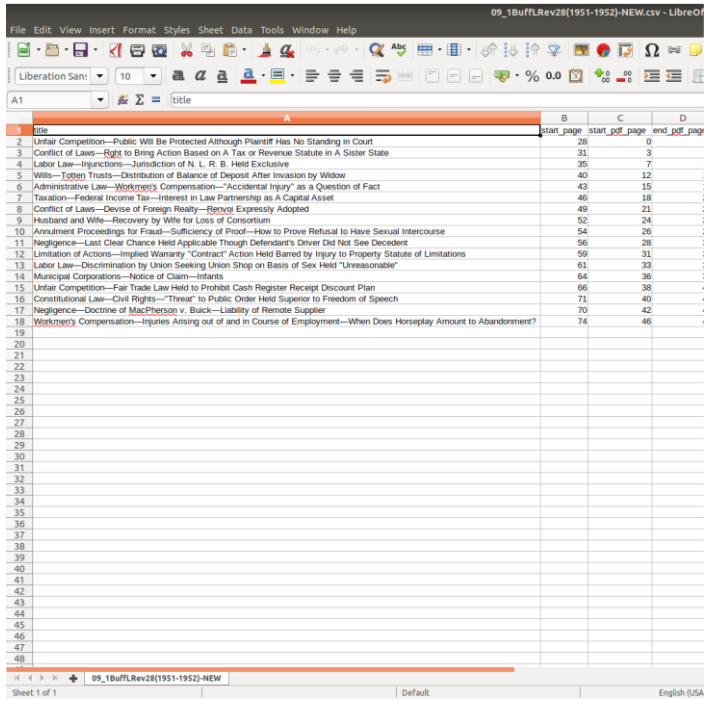
Feed hand-corrected CSV and original PDF back to dsplit-XX.py to split. For extra fun, hand-correct a couple of volumes, then use a bash script to run through them all while you get coffee.



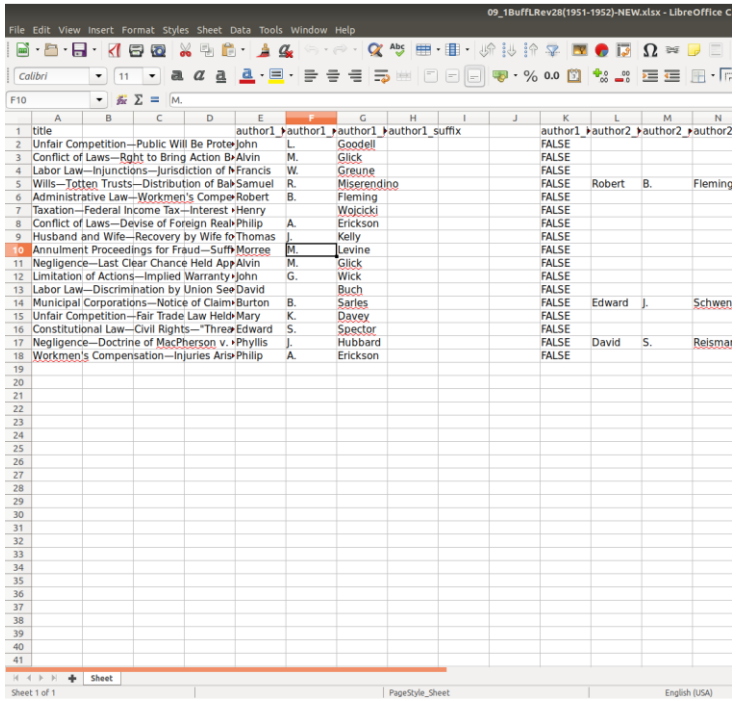
Convert CSV

Feed that CSV file to dc-convert.py. Copy everything back to the main computer. Cut and paste entries from exported Excel file into DC spreadsheet.

```
FileNotFoundError: [Errno 2] No such file or directory: '/media/john/B48A-17E4/Buffalo Law Review/burlf1/09_1*.csv'
john@john-UX31A:~/PycharmProjects/journaltools$ python3 dc-convert.py -v -p /media/john/B48A-17E4/buffalolawreview.xlsx '/media/joh
/B48A-17E4/Buffalo Law Review/buflr1/09_1'*.csv
Using template file /media/john/B48A-17E4/buffalolawreview.xlsx
Reading /media/john/B48A-17E4/Buffalo Law Review/buflr1/09_1BuffLRev28(1951-1952)-NEW.csv
Saving /media/john/B48A-17E4/Buffalo Law Review/buflr1/09_1BuffLRev28(1951-1952)-NEW.xlsx
john@john-UX31A:~/PycharmProjects/journaltools$
```



title	start_page	start_page_page	end_pdf_page
Unfair Competition—Public Will Be Protected Although Plaintiff Has No Standing in Court	28	0	
Conflict of Laws—Right to Bring Action Based on A Tax or Revenue Statute in A Sister State	31	3	
Labor Law—Injunctions—Jurisdiction of N. L. R. B. Held Exclusive	35	7	
Wills—Totten Trusts—Distribution of Balance of Deposit After Invasion by Widow	40	12	
Administrative Law—Workmen's Compensation—"Accidental Injury" as a Question of Fact	43	15	
Taxation—Federal Income Tax—Interest in Law Partnership as A Capital Asset	46	18	
Conflict of Laws—Devise of Foreign Realty—Renvoi Expressly Adopted	49	21	
Husband and Wife—Recovery by Wife for Loss of Consortium	52	24	
Annulment Proceedings for Fraud—Sufficiency of Proof—How to Prove Refusal to Have Sexual Intercourse	54	26	
Negligence—Last Clear Chance Held Applicable Though Defendant's Driver Did Not See Decedent	56	28	
Limitation of Actions—Implied Warranty "Contract" Action Held Barred by Injury to Property Statute of Limitations	59	31	
Labor Law—Discrimination by Union Seeking Union Shop on Basis of Sex Held "Unreasonable"	61	33	
Municipal Corporations—Notice of Claim—Infants	64	36	
Unfair Competition—Fair Trade Law Held to Prohibit Cash Register Receipt Discount Plan	66	38	
Constitutional Law—Civil Rights—"Threat" to Public Order Held Superior to Freedom of Speech	71	40	
Negligence—Doctrine of MacPherson v. Buick—Liability of Remote Supplier	70	42	
Workmen's Compensation—Injuries Arising out of and in Course of Employment—When Does Horseplay Amount to Abandonment?	74	46	

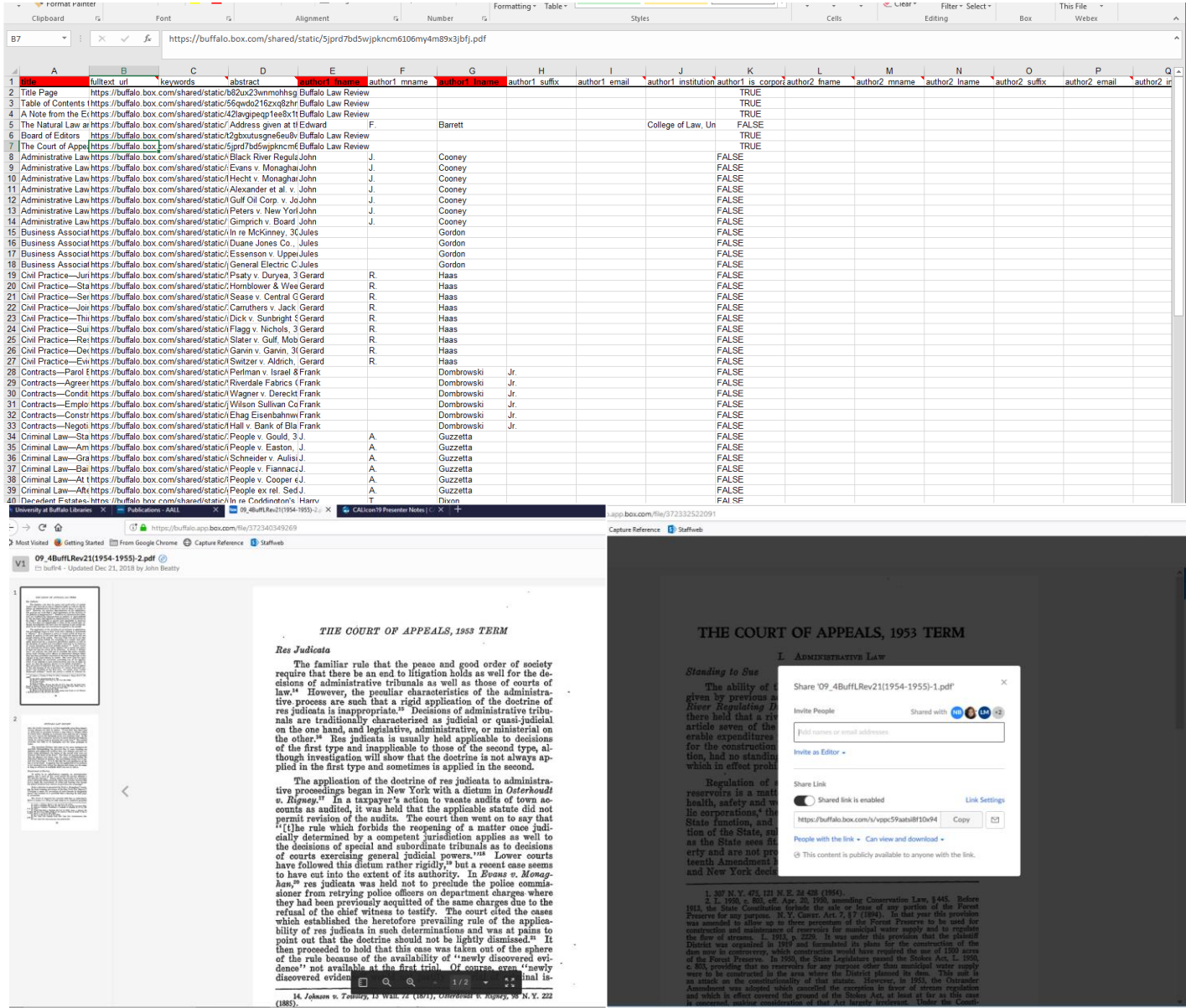


title	author1_	author1_	author1_	author1_suffix	author1_	author2_	author2_	author2
Unfair Competition—Public Will Be Prote	John	L.	Goodell		FALSE			
Conflict of Laws—Right to Bring Action B	Alvin	M.	Glick		FALSE			
Labor Law—Injunctions—Jurisdiction of N	Francis	W.	Greune		FALSE			
Wills—Totten Trusts—Distribution of Ba	Samuel	R.	Miserendino		FALSE	Robert	B.	Fleming
Administrative Law—Workmen's Compe	Robert	B.	Fleming		FALSE			
Taxation—Federal Income Tax—Interest	Henry		Wojcicki		FALSE			
Conflict of Laws—Devise of Foreign Real	Philip	A.	Erickson		FALSE			
Husband and Wife—Recovery by Wife fo	Thomas	J.	Kelly		FALSE			
Annulment Proceedings for Fraud—Suff	Morree	M.	Levine		FALSE			
Negligence—Last Clear Chance Held Ap	Alvin	M.	Glick		FALSE			
Limitation of Actions—Implied Warranty	John	G.	Wick		FALSE			
Labor Law—Discrimination by Union Se	David		Buch		FALSE			
Municipal Corporations—Notice of Claim	Burton	B.	Sarles		FALSE	Edward	J.	Schwen
Unfair Competition—Fair Trade Law Held	Mary	K.	Davey		FALSE			
Constitutional Law—Civil Rights—"Threa	Edward	S.	Spector		FALSE			
Negligence—Doctrine of MacPherson v.	Phyllis	J.	Hubbard		FALSE	David	S.	Reisma
Workmen's Compensation—Injuries Aris	Philip	A.	Erickson		FALSE			

University at Buffalo
School of Law

Hand-check as normal

Open split PDFs in Box preview. Check page split. Double-check metadata. Add disciplines. Cut and paste Box link.



journaltools.py

Main Python code—Contains all reusable code

- Author name and title manipulation (splitname, capitalize_title)
- PDF splitting code (splitpdf)
- PDF reading code (getpdf)
- CSV manipulation (importcsv, exportcsvnew, convertcsv)
- Page preparation (doublepages, croppages)
- PDF manipulation code (combinepdf, shiftpage, dirshift)
- Support code (getfilenames)

Most of these code segments called by external files that act as command-line interfaces

- E.g. dir-shift.py: Takes a path and passes it to dirshift

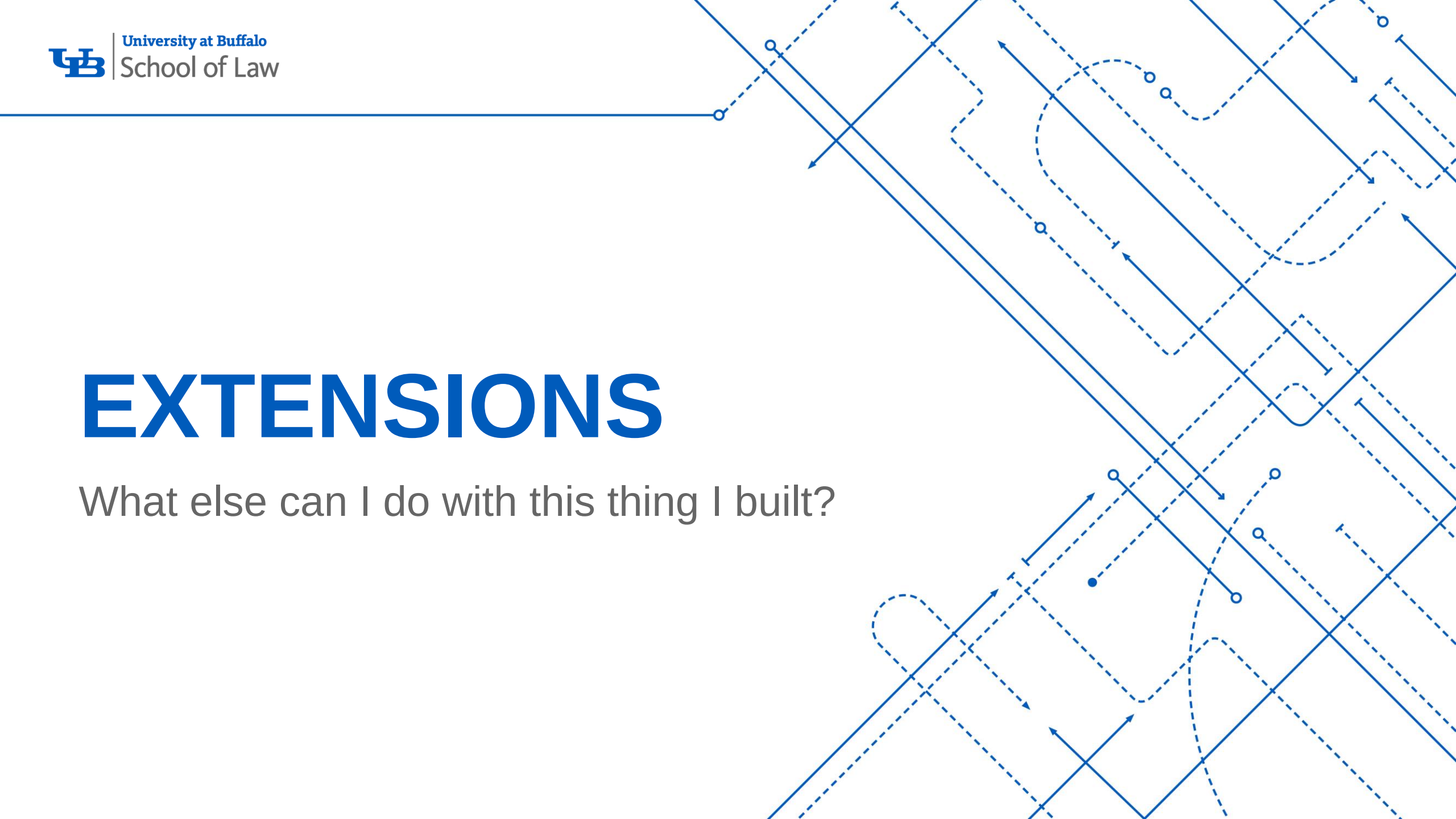
dsplit-XX.py

- This is the main metadata extraction and PDF splitting code.
- Different command line file is used for each type of file scanned
- Consists of a command-line interface and scanning code
- Remainder of code is the same for each. Calls to journaltools.py.

Other functions

- combine-pdf.py: Used to combine Hein-split volume indexes back into a single file. Takes a path and combines all files in filename order.
- dc-convert.py: Exports CSV file to an Excel file, with metadata in the proper columns to be cut and paste into DC upload sheets.
- dir-shift.py: Takes a path; copies the first page of every file and adds it as the last page of the previous file in the directory
- page-shift.py: Takes two files and copies the first page of the second file and adds it to the end of the first file (quickly replaced by dir-shift.py)

University at Buffalo
School of Law



EXTENSIONS

What else can I do with this thing I built?

New volumes of Buffalo Law Review

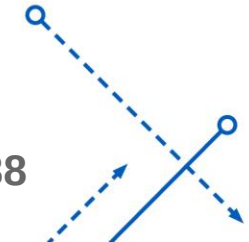
- Five new issues a year need to be processed and uploaded
- NO OCR text
- New command line program extracts metadata from a single file
- Bash script used to scan all articles and write to a single CSV
- Total processing time for an issue: About 15 minutes

UB Law Forum

- 38 volumes, 1-2 issues/volume
- OCR text too unpredictable to automatically scan for metadata
- Contents page fairly comprehensive
- Partial automation solution
- Contents text copied and pasted into text editor, cleaned up with search and replace, then copied into Excel file

UB Law Forum

- New code to crop from full magazine page scans to 8.5 x 11
- New code to convert hand-built Excel file to CSV
- PDF splitting and export command lines re-used

Resources

- The journaltools code, at GitHub
  https://github.com/johnrbeatty/journaltools
- These slides
  https://digitalcommons.law.buffalo.edu/law_librarian_other_scholarship/