Visualizing the Indianapolis Recorder: How Libraries can Promote Digital Scholarship

Ted Polley
Jenny Johnson
ILF, November 19, 2014

IUPUI

Indianapolis Recorder

-Background

-One of the top African-American publications in the nation

-Weekly publication began in 1899

-In 2009 IUPUI University received permission to digitize collection

IUPUI

Processing the collection

-Contracted with Lyrasis to digitize microfilm

-Use CONTENTdm to provide access

    -OCR extension

    -5202 issues, 90,461 of pages

Challenges with Microfilm

-Poor Image Quality

THE RECORDER, INDIANAPOLIS, INDIANA

"BILLY" AND "TEDDY"

BRYAN AND ROOSEVELT MEET BY CHANCE.

WRECK ON MONON.

FLOODS IN MEXICO.

STRIKE AT WEST POINT.

DRAMATIC SCENES

YOUTSEY BECOMES HYSTERICAL IN COURT ROOM.

ANDREE'S FATE.

WHAT THEY SAY

VARIOUS VIEWS BY DIVERS POLITICAL PROPHETS.

THRIFTY GLASS CUTTERS

ENGLISH ELECTIONS.

Do You Know

THAT

THE RECORDER

Is Prepared to do all kinds of Job Printing on short notice? We can make anything from a Bill Head, Letter Head, Minutes, Dodgers, Tickets, Business Cards, Visiting Cards, Book or Newspaper,

In fact, everything in Job Printing

We make a specialty of first-class Job Printing

If you want anything done and are too busy to call, drop us a postal or call telephone 561.

If You Want any kind of Printing done let us know.

If you have anything to advertise send it to The Recorder.

If you have a house to rent,

If you have a room to let,

If you want a situation,

If you want to sell anything,

If you want anything,

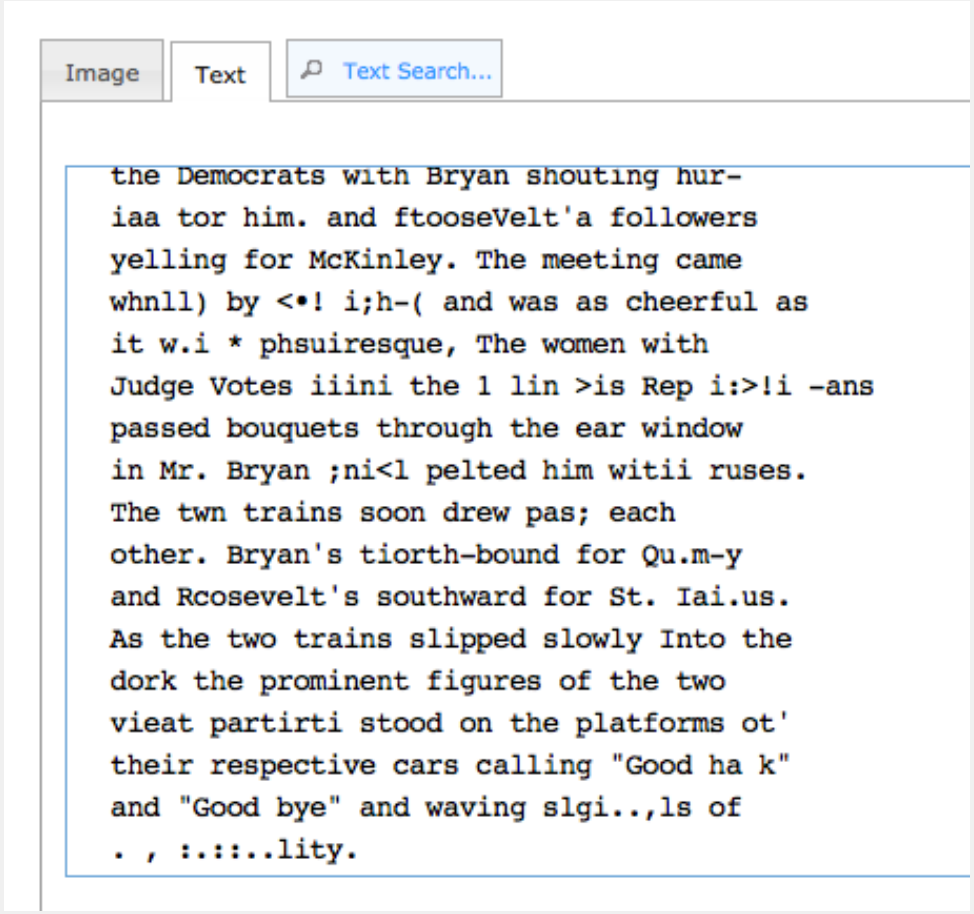
Advertise in

THE RECORDER

We Want your subscription, your advertisement.

Challenges with Microfilm

Issues with OCR



IUPUI

Collection Inquiries

-Basic questions from general public about how to use CONTENTdm

    -basic browsing and searching

    -Printing specific articles

-In recent years there has been an increase in requests from scholars and researchers

    -Interest in text mining

IUPUI

Potential Scholarly Uses for the Recorder

This resource is potentially valuable to researchers studying:

- The History of Indianapolis' African American community
- The evolution of language over time
- The changes in content and structure of a newspaper to meet changing community needs and interests

IUPUI

Supporting Digital Scholarship

Digital humanities include a broad range of scholarly activities drawing from the field of computing and disciplines in the humanities.

Projects can include the curation of cultural datasets in digital collections, analysis of large text corpora, social network analysis, and data visualization.

As part of its mission, the University Library Center for Digital Scholarship provides support and consultation to IUPUI scholars engaged in, or planning digital humanities projects.

IUPUI

Promoting the Recorder to IUPUI Researchers

To promote this valuable resource to the rest of IUPUI's campus the Center for Digital Scholarship conducted an exploratory study using basic approaches to text mining and visualization.

We focused on characterizing the semantic content of the Recorder, creating topic maps to facilitate exploration of this resource.

Selecting a Sample

The full text of the newspaper with metadata can be downloaded in either XML or tab-delimited format.

The first (1899) and the last (2005) years in the collection were selected as a sample.

The transcripts for every issue for those years were extracted from the dataset, metadata were stripped, and the issues from each year were combined and saved as two text files.

Generating the Visualizations

The transcripts for each year were visualized using VOSviewer, a free program for analyzing text and generating topic maps.

The program calculates term frequency, how often terms co-occur, and statistical distributions of terms to determine their relevancy in the overall corpus.

Van Eck & Waltman, 2011

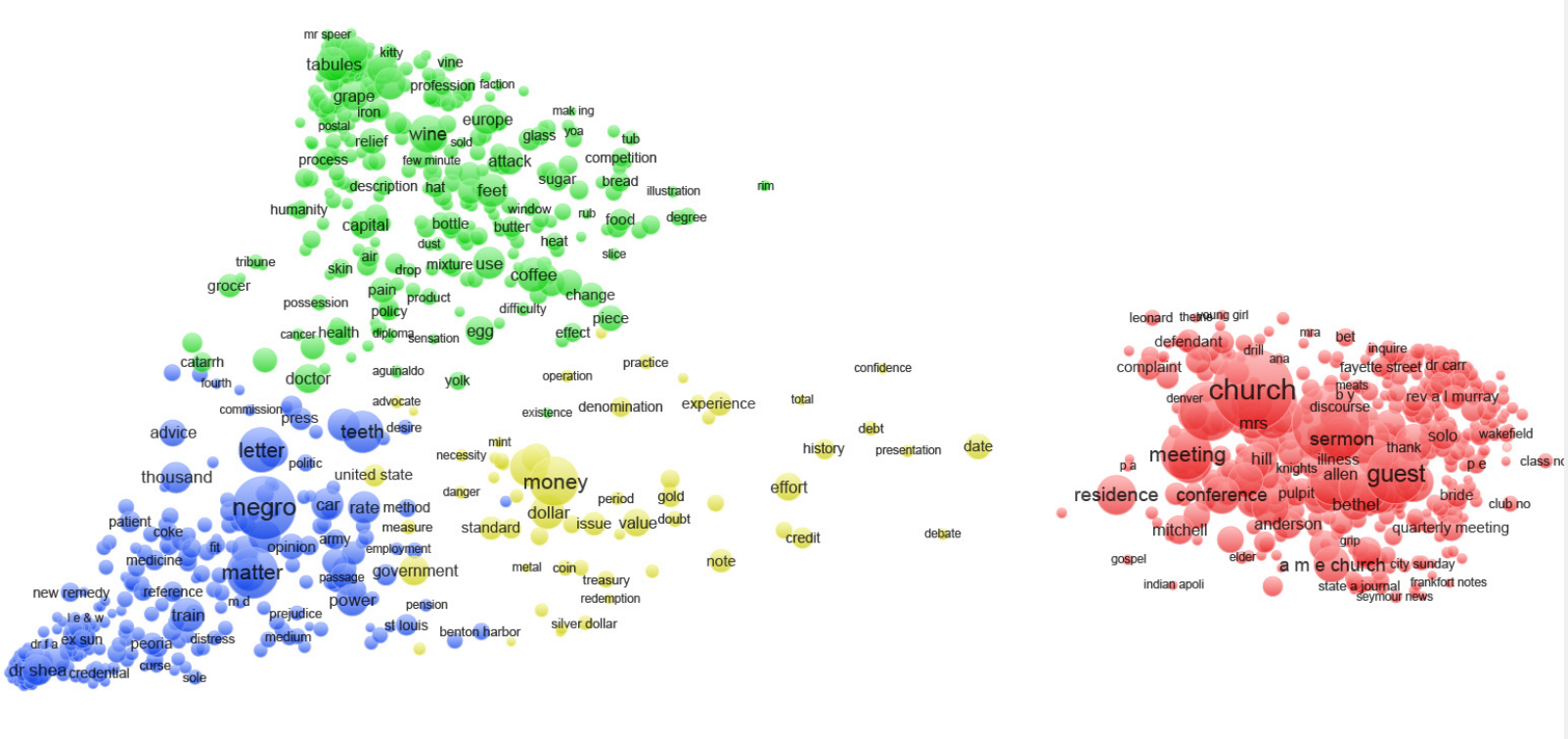
1899 Top 10 Most Frequently Occurring Terms

Term	Occurrences	Relevance
church	802	0.66
pastor	745	0.84
negro	500	0.62
club	433	0.61
guest	423	0.93
matter	328	0.74
meeting	328	0.57
smith	304	0.87
money	295	0.28
johnson	284	0.84

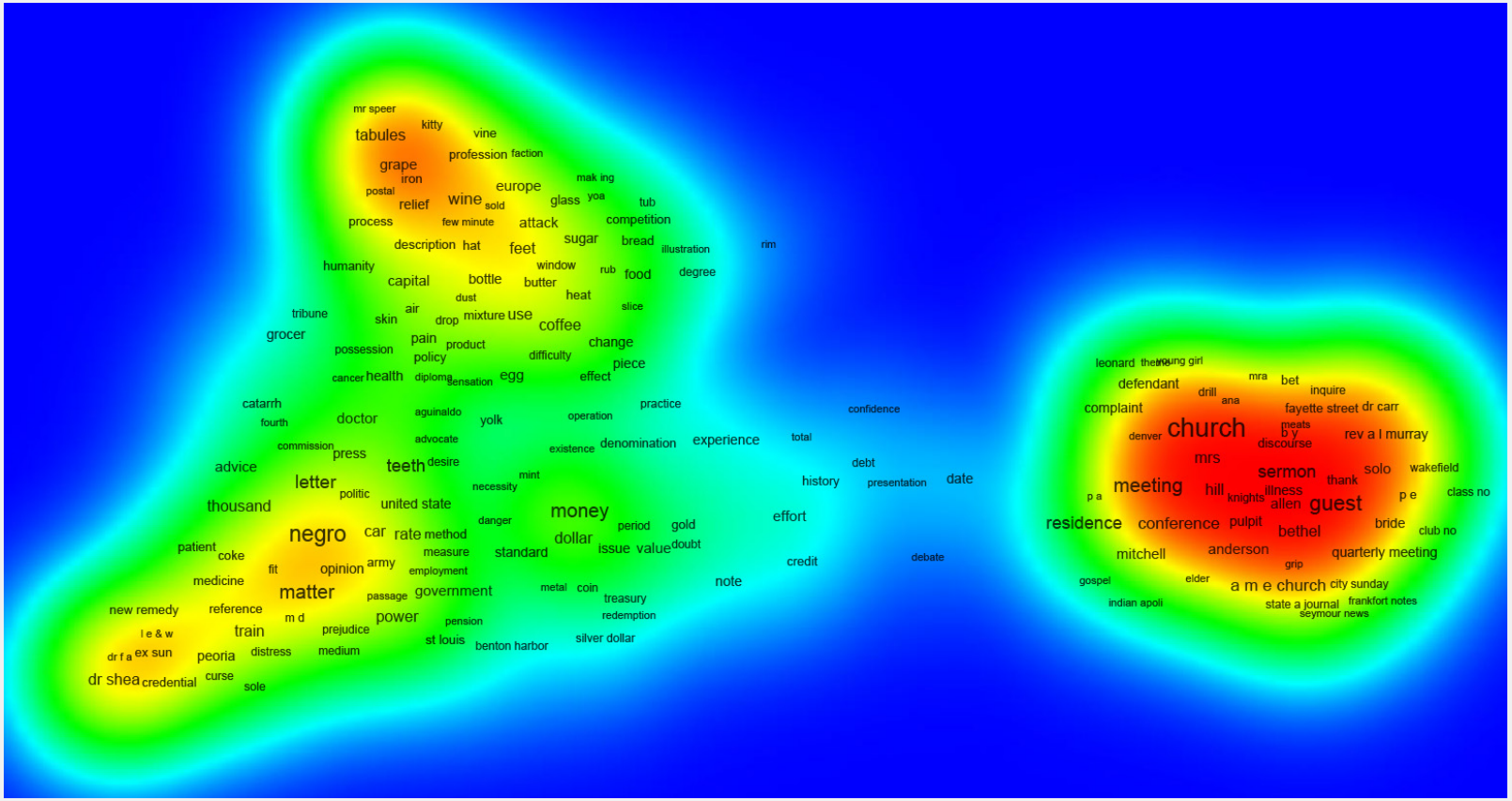
1899 Top 10 Most Relevant Terms

Term	Occurrences	Relevance
gold standard	13	2.99
silver certificate	20	2.96
silver dollar	25	2.88
greenback	10	2.78
certificate	23	2.73
redemption	13	2.64
coin	32	2.59
metal	11	2.51
silver	31	2.49
statute	21	2.33

1899 Topic Map



1899 Topic Map



2005 Top 10 Most Frequently Occurring Terms

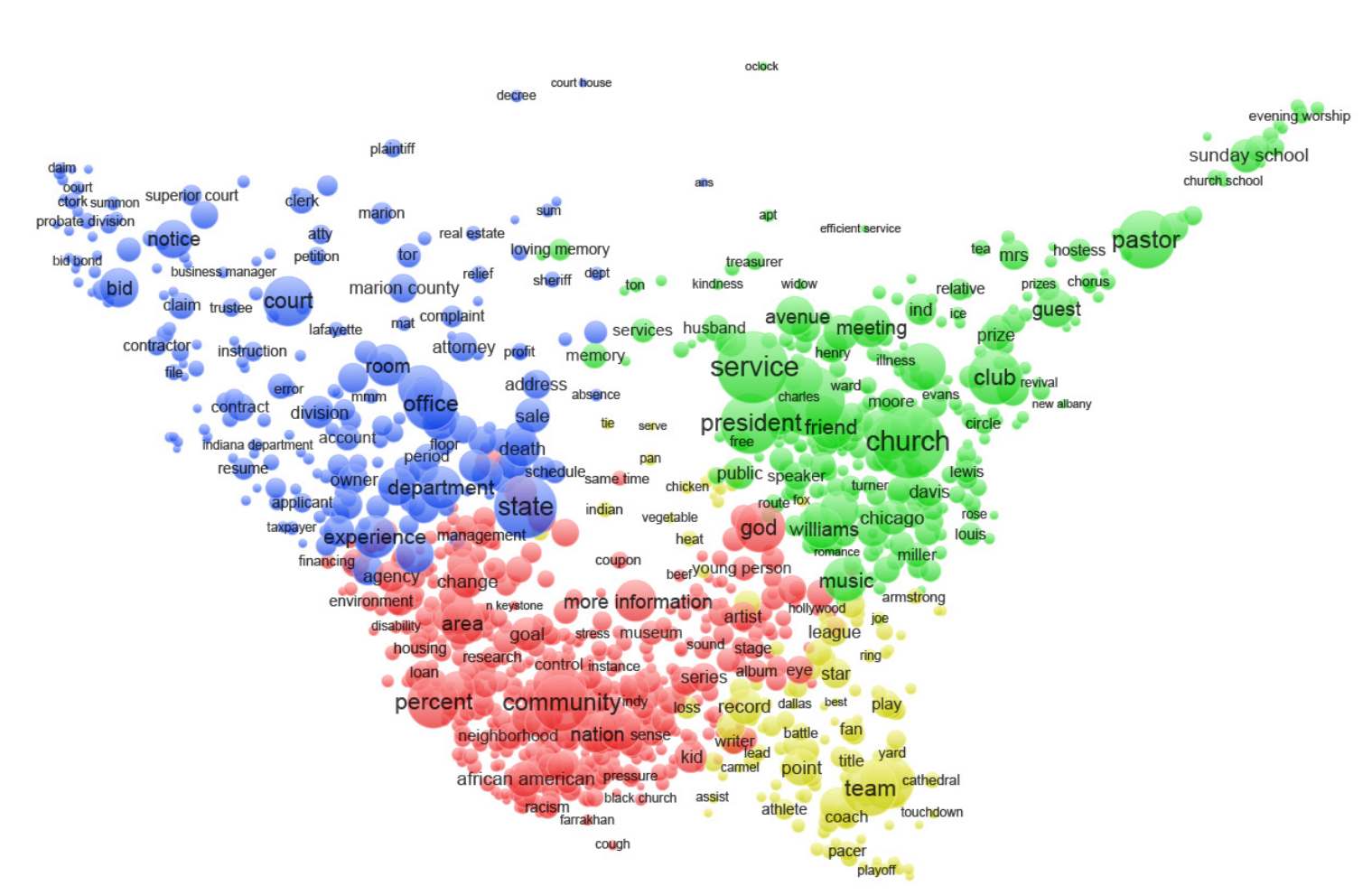
Term	Occurrences	Relevance
church	13721	0.76
service	13071	0.69
home	12964	0.43
state	9873	0.29
president	9376	0.30
pastor	8461	3.16
community	8199	0.44
team	7854	1.74
game	7440	1.60
club	7435	0.94

IUPUI

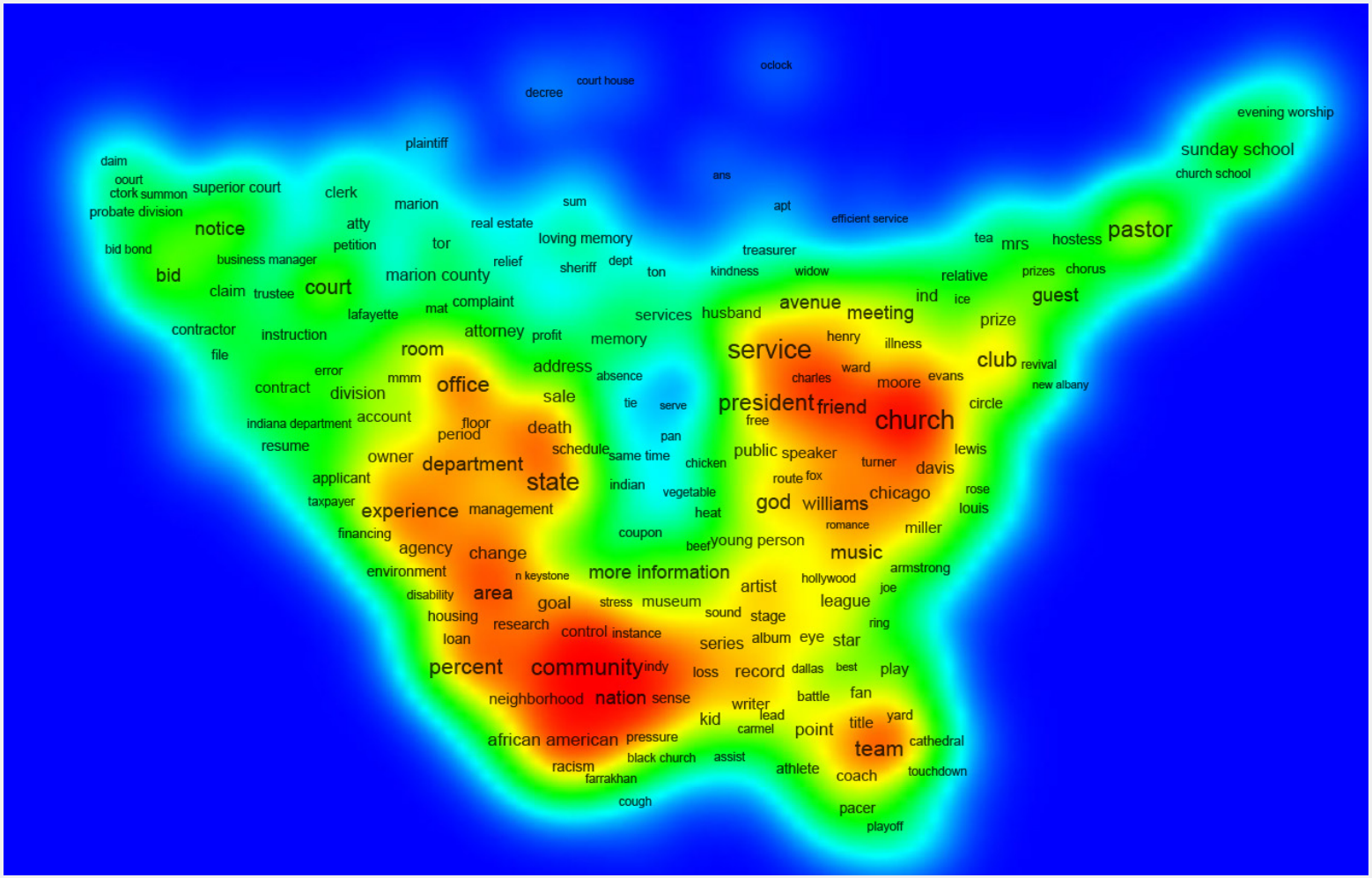
2005 Top 10 Most Relevant Terms

Term	Occurrences	Relevance
regular service	494	5.80
pastor Sunday school	185	5.76
evening worship	549	5.67
prayer meeting	288	5.57
bypu	447	5.57
p u	459	5.57
service Sunday school	243	5.52
marion suparior [*sic*] court	197	5.33
morning worship	1615	5.31
sundayschool	338	5.24

2005 Topic Map



2005 Topic Map



Goals for OCR Improvement

Improve information retrieval

Increase the utility of the text corpus as data

- Text mining
- Cross-linking information about people, places, or organizations from multiple sources
- Other types of digital humanities research

Where does this leave those of us managing digital collections of historical newspapers?

Options for Improvement

Scan from paper copies (instead of microfilm)
Manual correction
Train the OCR engine
Create a custom dictionary

Holley, 2008
Arlitsch & Herbert, 2004

IUPUI

Decision Making Process

~~Scan from paper copies (instead of microfilm)~~
- Not feasible to get print copies to re-digitize

~~Manual correction~~
- Manual correction is extremely time-consuming

~~Train the OCR engine~~
- No ability to train the OCR engine within ContentDM

Create a custom dictionary
- Possible, if we could over some a few challenges

Custom Dictionary (of what?)

Proper nouns – names and place names are harder for OCR engines to process (Tanner et al, 2009)

After scanning the front pages of several issues, we decided to create custom dictionaries for Indianapolis and Central Indiana centric:

- People
- Places
- Organizations

Experiment: Custom Dictionary

Challenges

- OCR software (ABBYY FineReader) is run on the server
- No ability to import custom dictionaries

Work-arounds (for this experiment)

- Purchase a local installation of ABBYY FineReader 12
- Run OCR on the sample manually

Experiment Details

Considerations

- Relatively low time investment (manually running OCR by issue is slow) compared to other approaches
- Potential to improve retrieval of proper names most significant to Indianapolis Recorder readers and local historians

Sample:

- 6 issues randomly selected each year (1960 – 1969)
- 60 issues (+ 2 special issues) in total
- Generated a random sample in Excel (RAND function) across the 10-yr timeframe

Process

Creating the dictionaries

- Manually read front pages of selected issues
- Enter people, place, and organization names used into Excel
- Converted excel tabs into text files
- Copied text files into designated ABBYY dictionary folder

Pre & Post OCR

- Local installation of ABBYY FineReader 12
- Pre – OCR was run using default installation and settings
- Post – OCR was run with dictionaries in place

Results

Both pre and post OCR files were combined into two text files.

These two files were then visualized using VOSviewer to determine if OCR correction makes a difference in resulting visualizations.

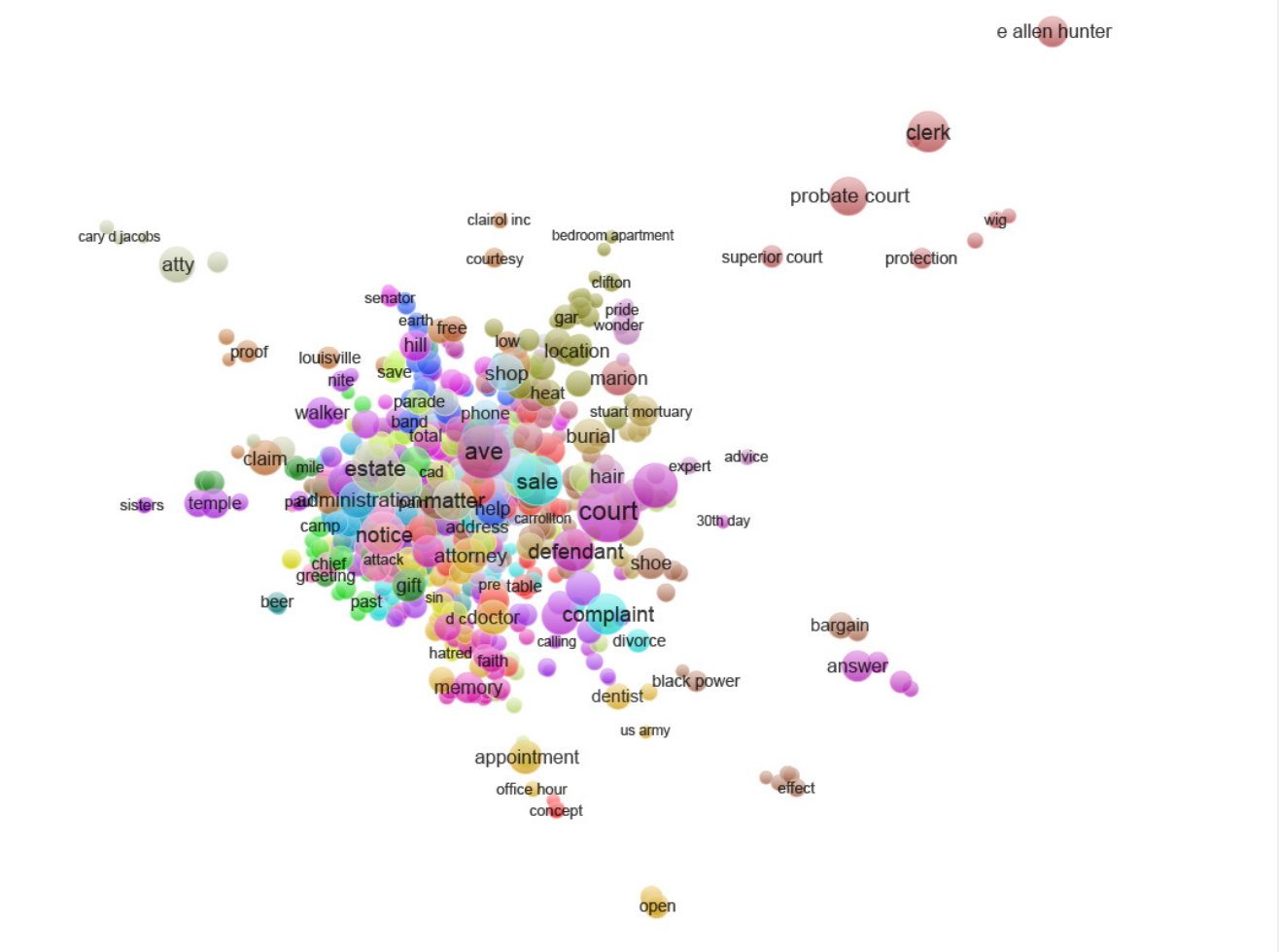
Pre OCR Top 10 Most Frequently Occurring Terms

Term	Occurrences	Relevance
court	266	0.60
ave	196	0.63
estate	171	0.72
sale	168	0.62
funeral service	138	0.85
marion county	138	0.69
notice	136	0.90
defendant	122	0.68
complaint	116	0.82
clerk	115	0.90

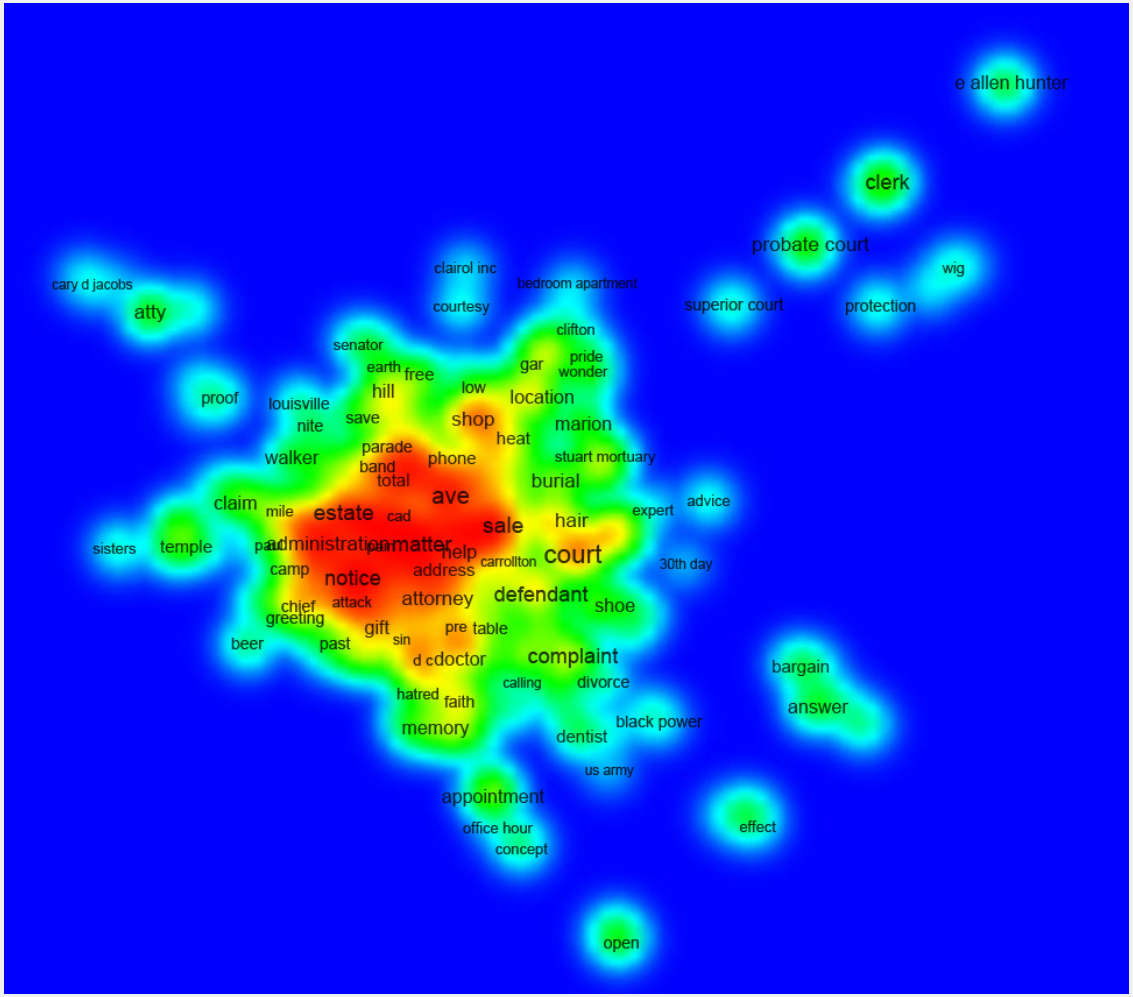
Pre OCR Top 10 Most Relevant Terms

Term	Occurrences	Relevance
absence	29	2.13
basileus	12	2.13
chevrolet	16	2.13
controversy	16	2.13
eagle	22	2.13
james burres	28	2.13
middle	16	2.13
pta	11	2.13
school system	23	2.13
ship	17	2.13

Pre OCR Topic Map



Pre OCR Topic Map



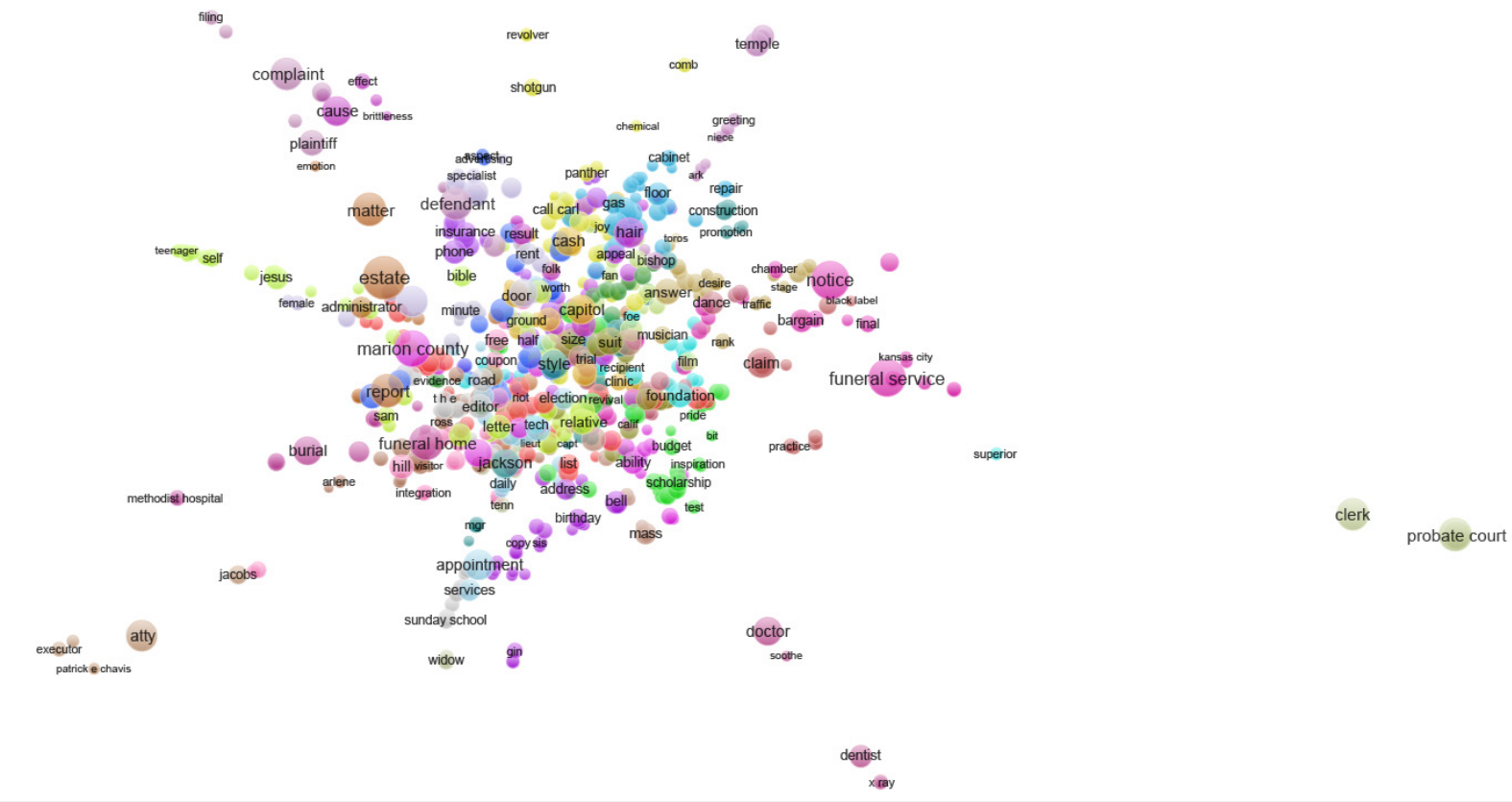
Post OCR Top 10 Most Frequently Occurring Terms

Term	Occurrences	Relevance
estate	197	0.74
notice	160	0.92
funeral service	145	1.18
marion county	143	1.04
funeral home	125	1.28
report	121	0.75
probate court	118	0.95
matter	118	0.74
defendant	114	0.76
clerk	108	0.94

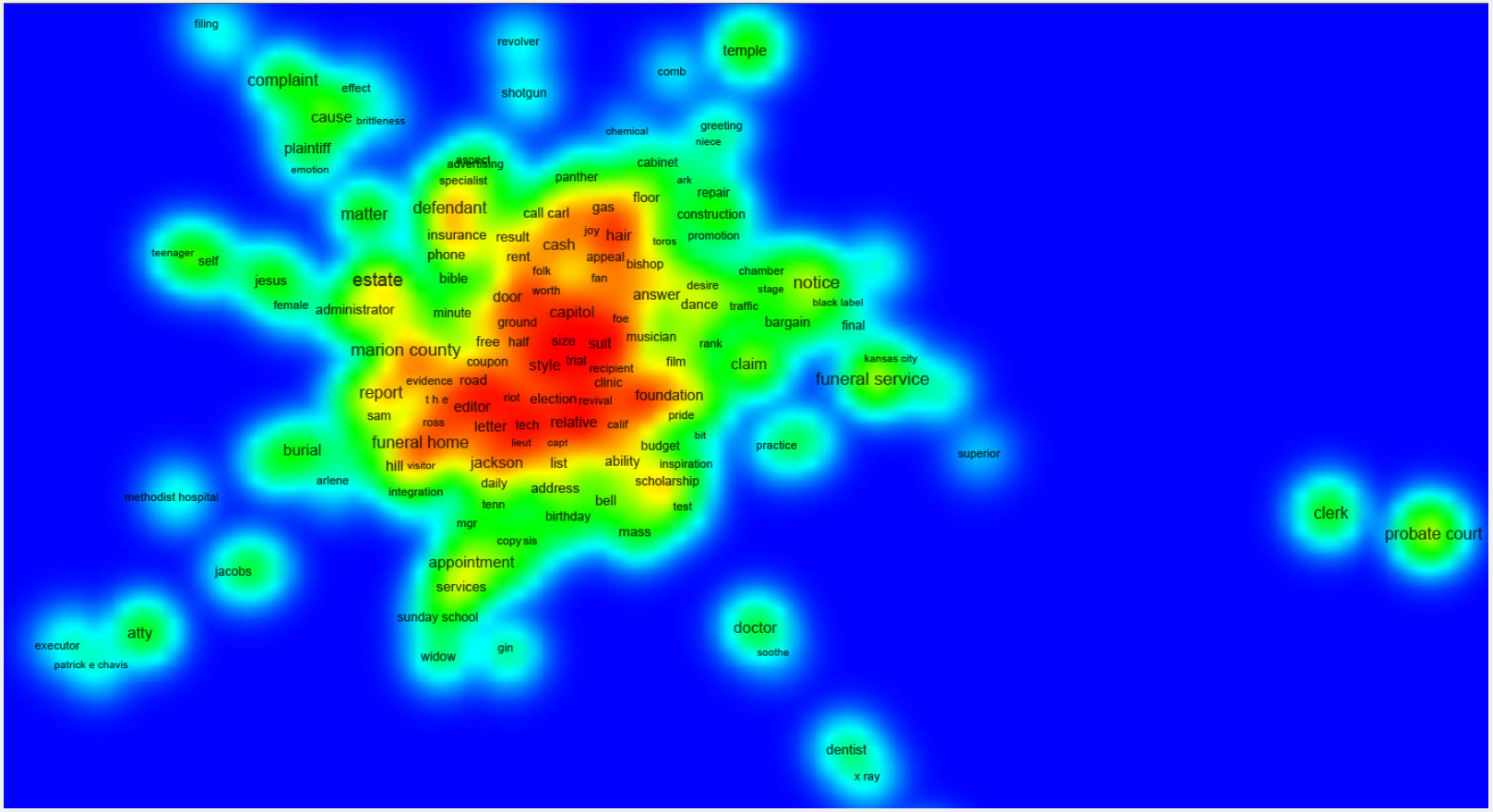
Post OCR Top 10 Most Relevant Terms

Term	Occurrences	Relevance
absence	28	1.98
chevrolet	17	1.98
comment	13	1.98
disciple	15	1.98
eagle	25	1.98
extension	19	1.98
highway	13	1.98
lafayett	15	1.98
lafayett square	13	1.98
n central	11	1.98

Pre OCR Topic Map



Pre OCR Topic Map



Acknowledgements

Thanks to Lisa Calvert for her help in creating the dictionaries.

Thanks to Anna Proctor and Lucy Williams for their input and thoughtful suggestions during the planning phase of this project.

References

1. Allen, R. B., Waldstein, I., & Zhu, W. (2008). Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres. *Lecture Notes in Computer Science,* 5362, 379-386.
2. Holley, R. (2009). How good can it get?: Analysing and improving OCR accuracy in large scale Historic Newspaper Digitisation Programs. *D-lib Magazine,* 15(3-4).
3. Klijn, E. (2008). The current state of art in newspaper digitisation. A market perspective. *D-Lib Magazine*, 14(1/2). doi: 10.1045/january2008-klijn
4. Strange, C., McNamara, D., Wodak, J., & Wood, I. (2014). Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8(1).
5. Van Eck, N.J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. ISSI Newsletter, 7(3), 50-54 .