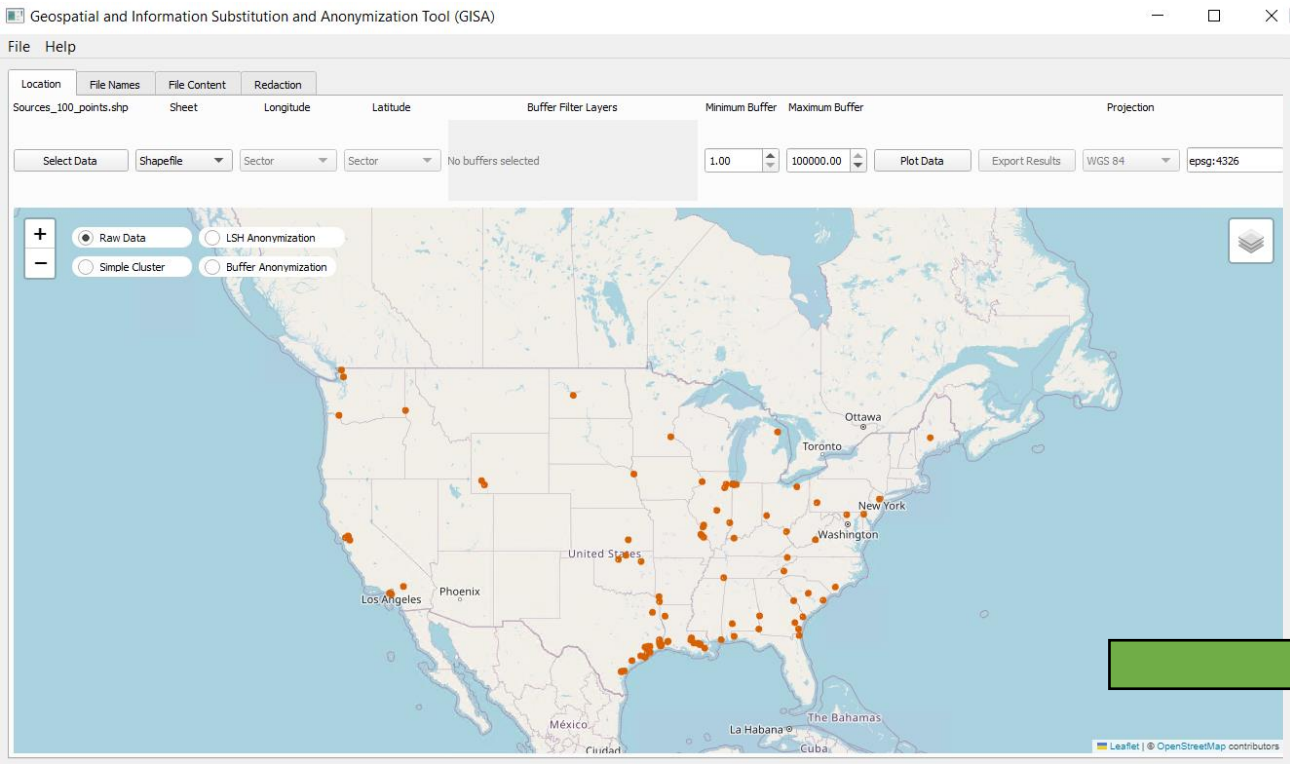
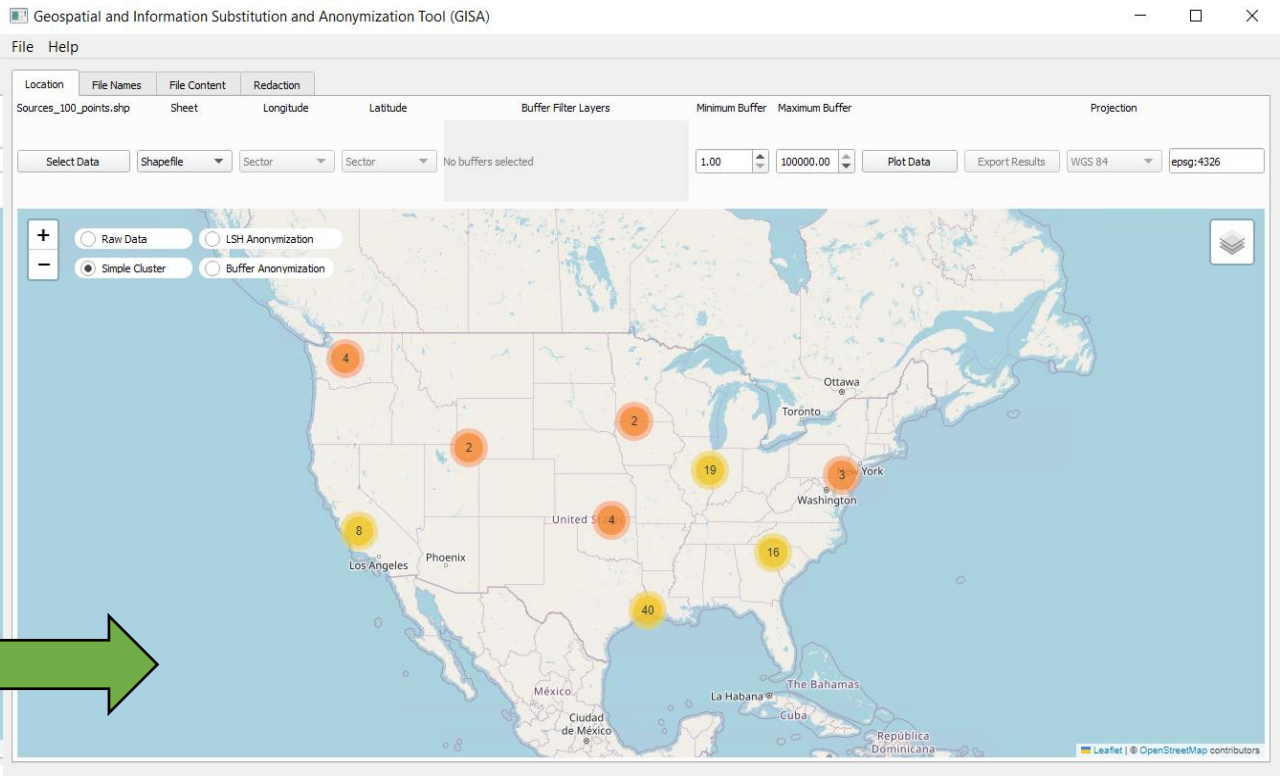
GISA- Protecting Stakeholder Spatial Data through an Advanced Anonymization Method

Patrick Wingo
Federal Research Scientist





FECM/NETL Carbon Management Research Project Review Meeting

Aug. 5, 2024

Disclaimer

This project was funded by the United States Department of Energy, National Energy Technology Laboratory, in part, through a site support contract. Neither the United States Government nor any agency thereof, nor any of their employees, nor the support contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Acknowledgement: This work was performed in support of the U.S. Department of Energy's (DOE) Office of Fossil Energy and Carbon Management's Geo-Analysis and Monitoring Team and was developed jointly through the U.S. DOE Office of Fossil Energy and Carbon Management's EDX4CCS Project, in part, from the Bipartisan Infrastructure Law.

Authors and Contact Information

Patrick Wingo[1], Paige Morkner[1], Michael Gao[1,2]

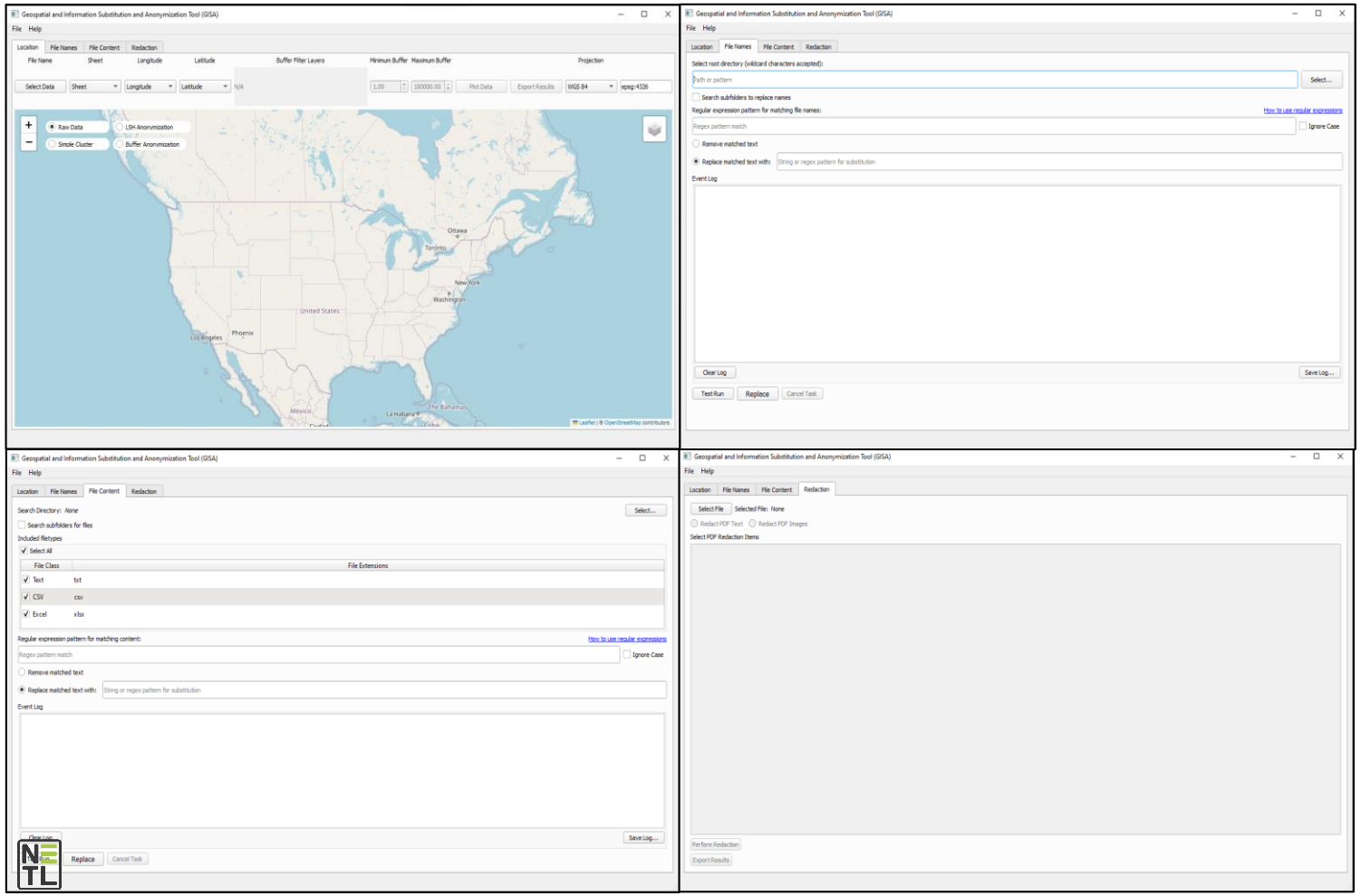
[1]National Energy Technology Laboratory, 1450 Queen Avenue SW, Albany, OR 97321, USA

[2]NETL Support Contractor, 1450 Queen Avenue SW, Albany, OR 97321, USA

Overview

Tour of GISA tool

- Founding Challenge & Objective
- Overview of Tool Functionality
  - **Geospatial Anonymization**
  - File Name/Content Substitution Anonymization
  - Redaction Anonymization
- Lessons Learned
- Next Steps
- Acknowledgements



Founding Challenge & Objective

The Geospatial and Information Substitution and Anonymization Tool (GISA)

Challenge:

- Data shared by industry partners frequently contains sensitive information
- Sensitive information can prevent or significantly delay sharing with other entities, public
- Removing sensitive information via *anonymization* allows for derivatives to be shared
- Anonymization of large, heterogeneous datasets can be time consuming

Objective:

- Create and deploy a tool to aid with anonymization of various types of data
- Provide multiple approaches, including:
    - Spatial Relocation
    - Substitution
    - Redaction

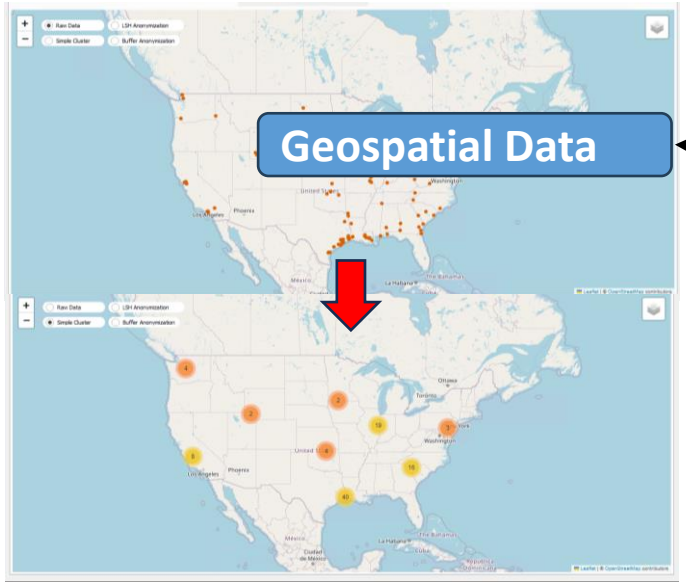
Tool: GISA - Geospatial Information Substitution & Anonymization

- Developed under EDX4CCS Task 46 (POP: August 2022 - March 2024)
- Desktop-based, available publically on EDX

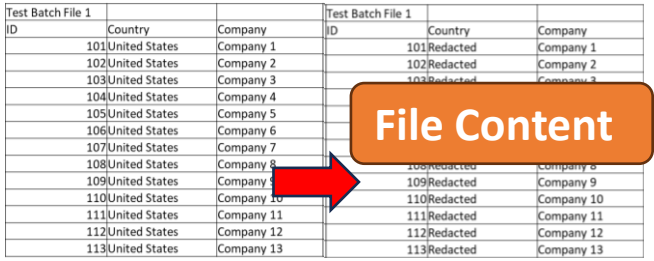
Overview of Functionality

Different Approaches to Anonymization

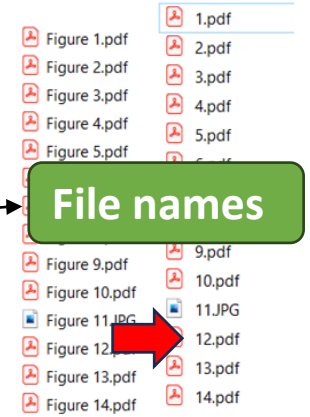
Rearrange points into approximate coordinates *without* revealing true location



Change text across many files



Tool to Anonymize:



Bulk Rename files following find-replace patterns

Apply redaction to specified text and images in PDFs

Geospatial Anonymization

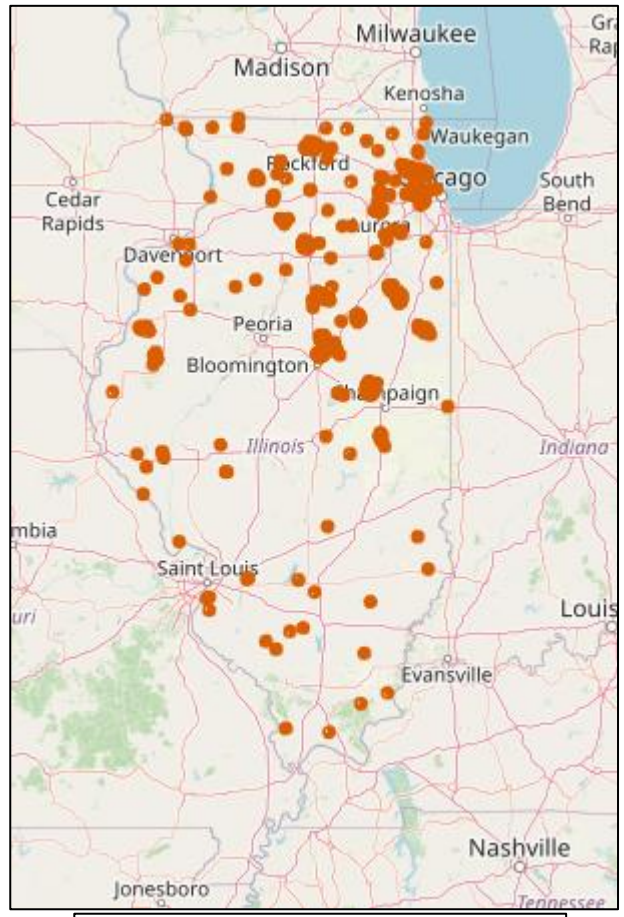
Adjust Points to approximate coordinates

Spatial coordinates can be sensitive, but we may still want to run analyses on a collection of points.

- Shuffle points without completely destroying relative/neighborhood relationships.
- Different approaches have different advantages, drawbacks.

GISA provides three methods of anonymization:

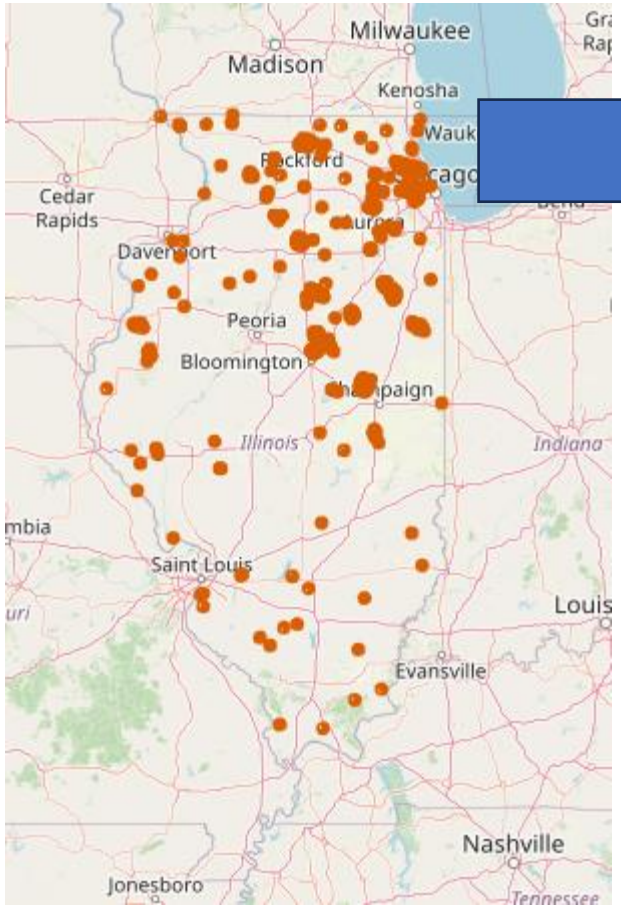
- Clustering
- Locality Sensitive Hashing (LSH)
- Constrained buffer



Example Point Dataset

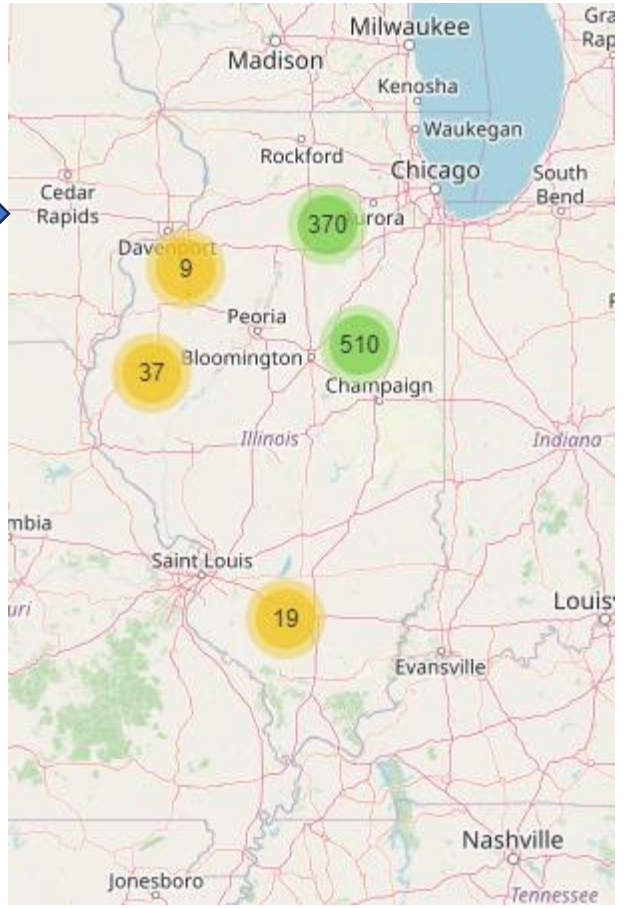
Geospatial Anonymization

Cluster Anonymization



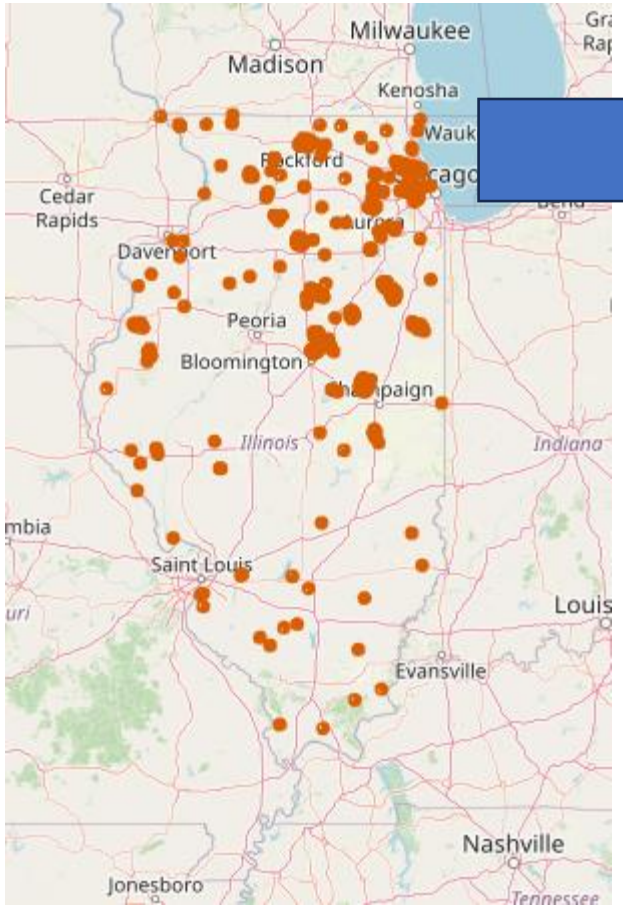
Apply Clustering

- Updates Clusters Dynamically based on zoom level
- Dynamic Visualization only (no export)



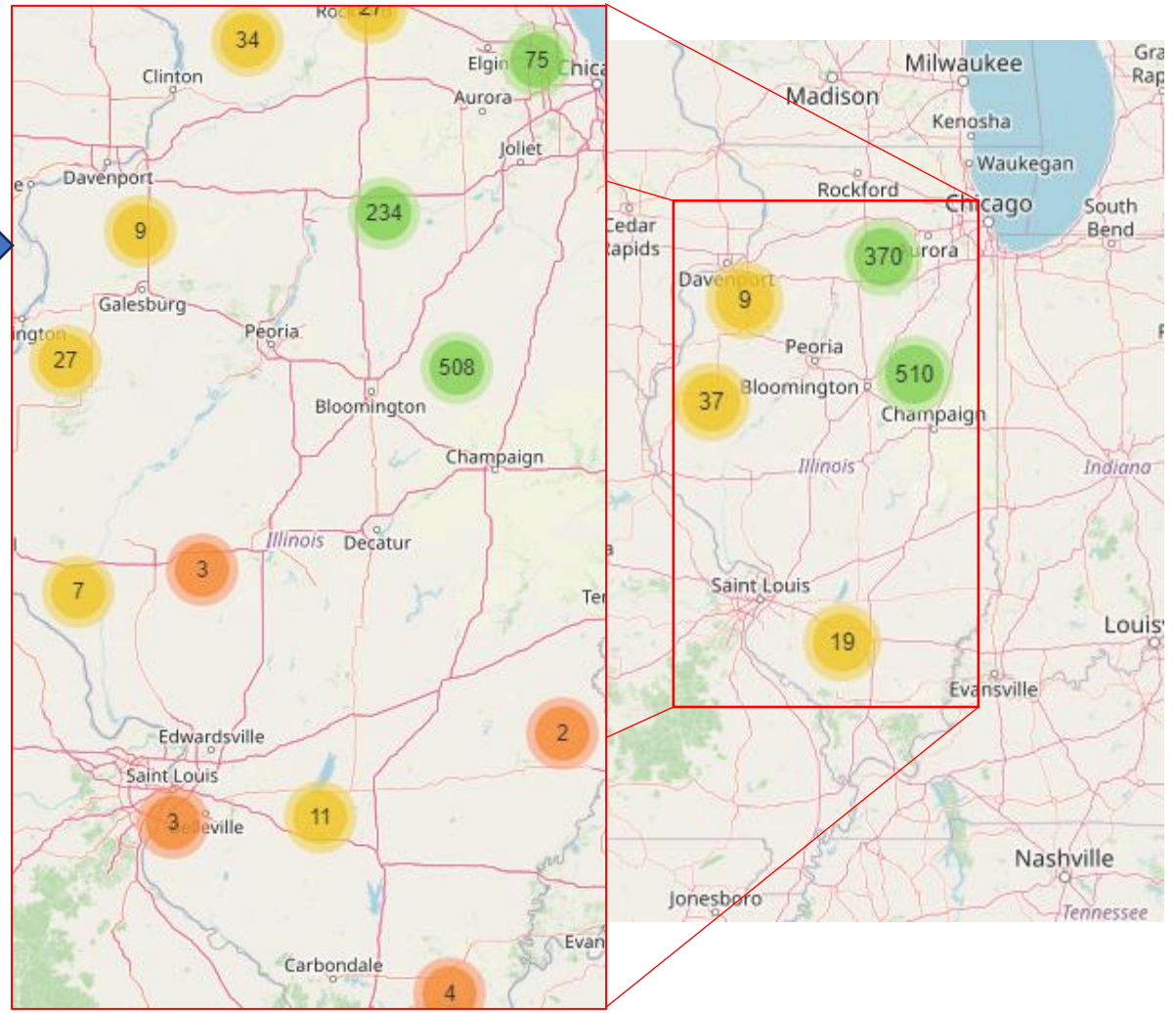
Geospatial Anonymization

Cluster Anonymization



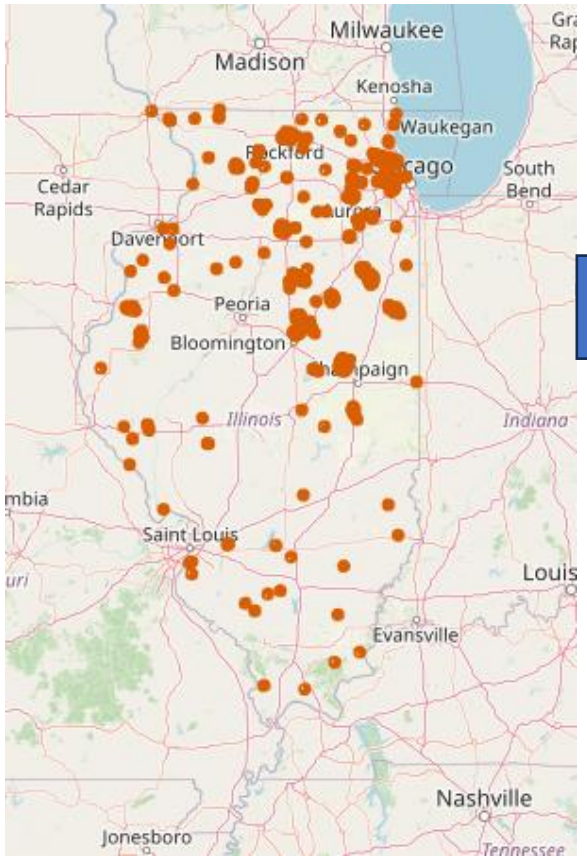
Apply Clustering

- Updates Clusters Dynamically based on zoom level
- Dynamic Visualization only (no export)



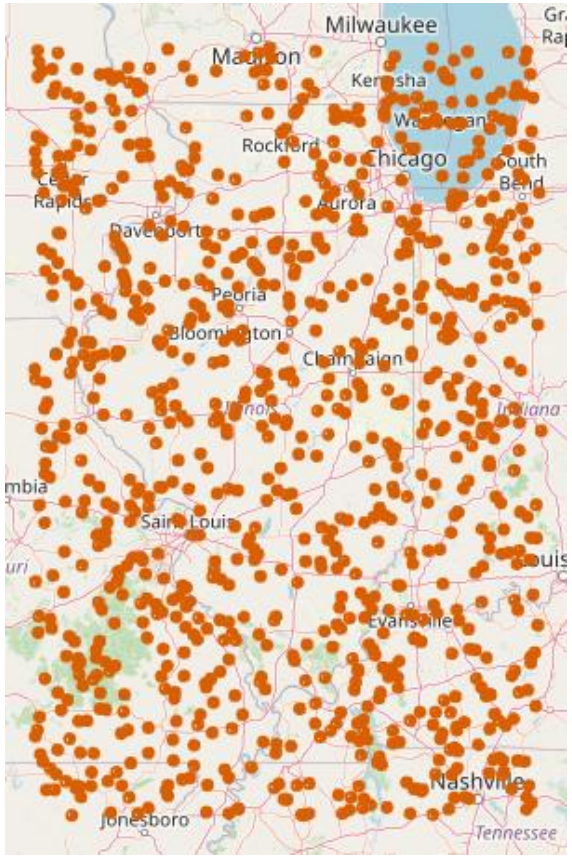
Geospatial Anonymization

Locality Sensitive Hashing (LSH) Anonymization



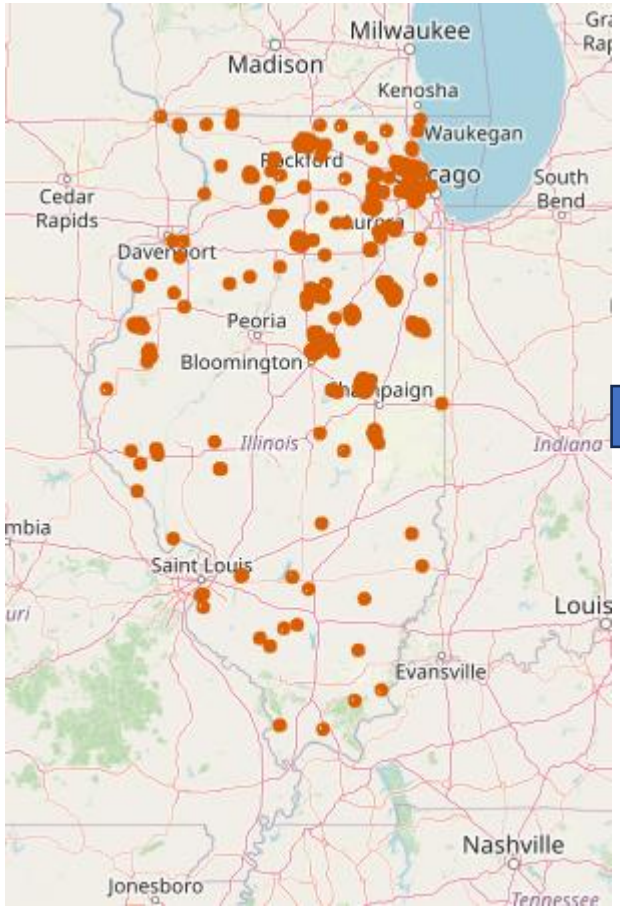
Apply LSH

- Piecewise Randomizes positions within extent of data
- Keeps points in region, impossible to reconstruct original location
- Destroys most proximity information

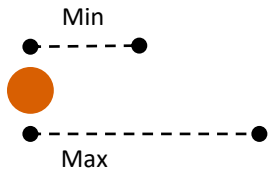


Geospatial Anonymization

Constrained Buffer Anonymization

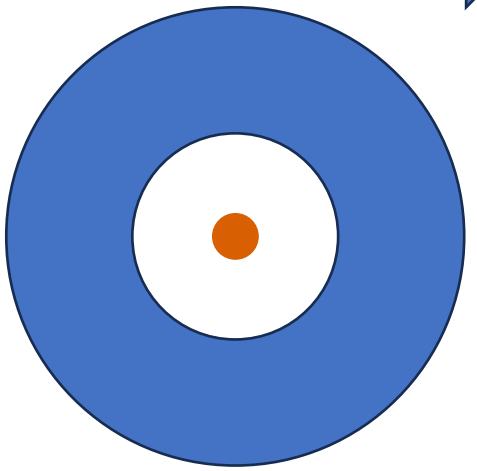


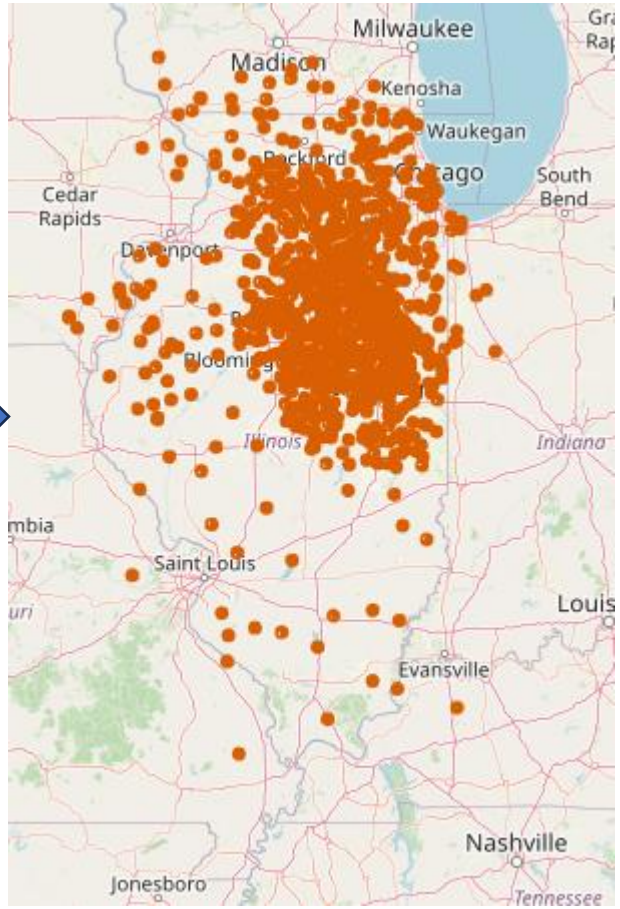
For each point, a minimum and maximum distance are applied



Apply Buffers and constraints

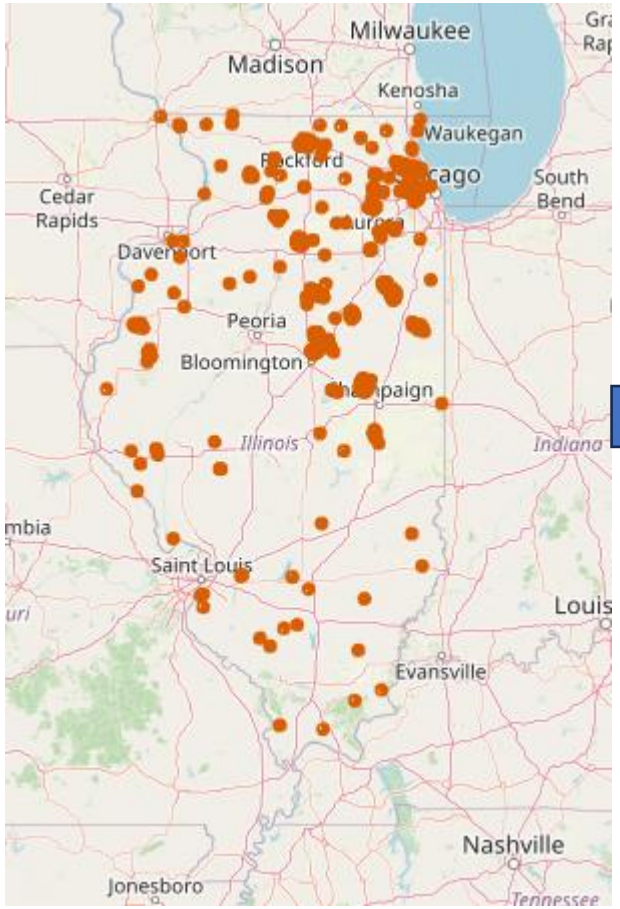
A “donut” shaped resampling region is produced



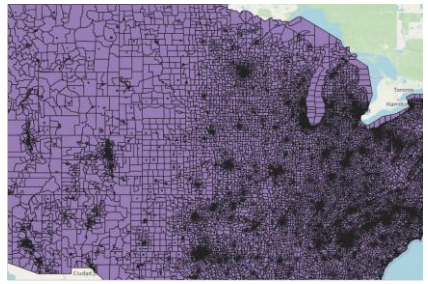


Geospatial Anonymization

Constrained Buffer Anonymization

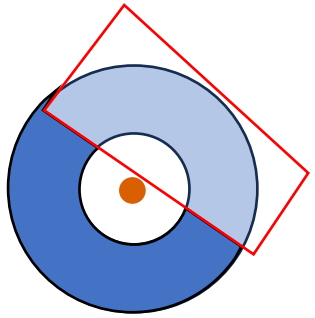


- Each Point constrained to overlapping polygons in “constraint layers”



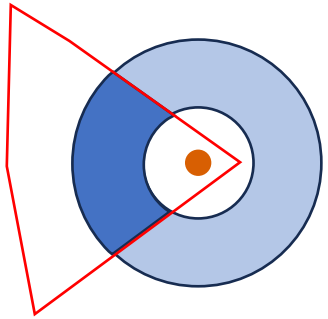
Apply Buffers and constraints

Point outside polygon?

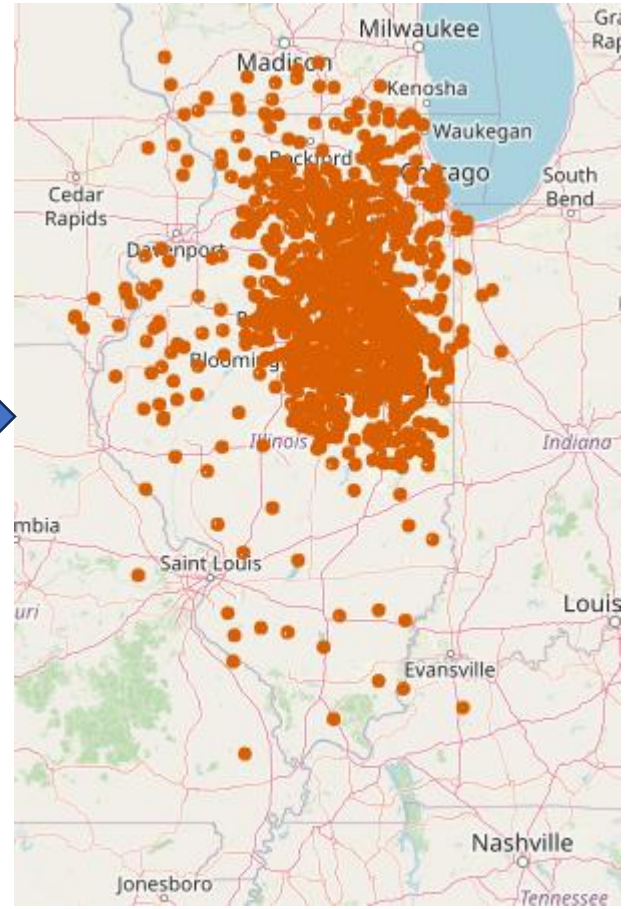


Sample buffer outside of polygon

Point inside polygon?



Sample buffer inside of polygon



File Name/Content Substitution Anonymization

Overview

Features:

- **User specifies terms** to **remove or replace** from batch **file names or file content**
- Utilize Regular Expressions (RegEx) to match strings (file names or file content)
- Remove or replace matches (can use Capture Group substitution)
- Can perform test run prior to actual changes
- Full log captures changes

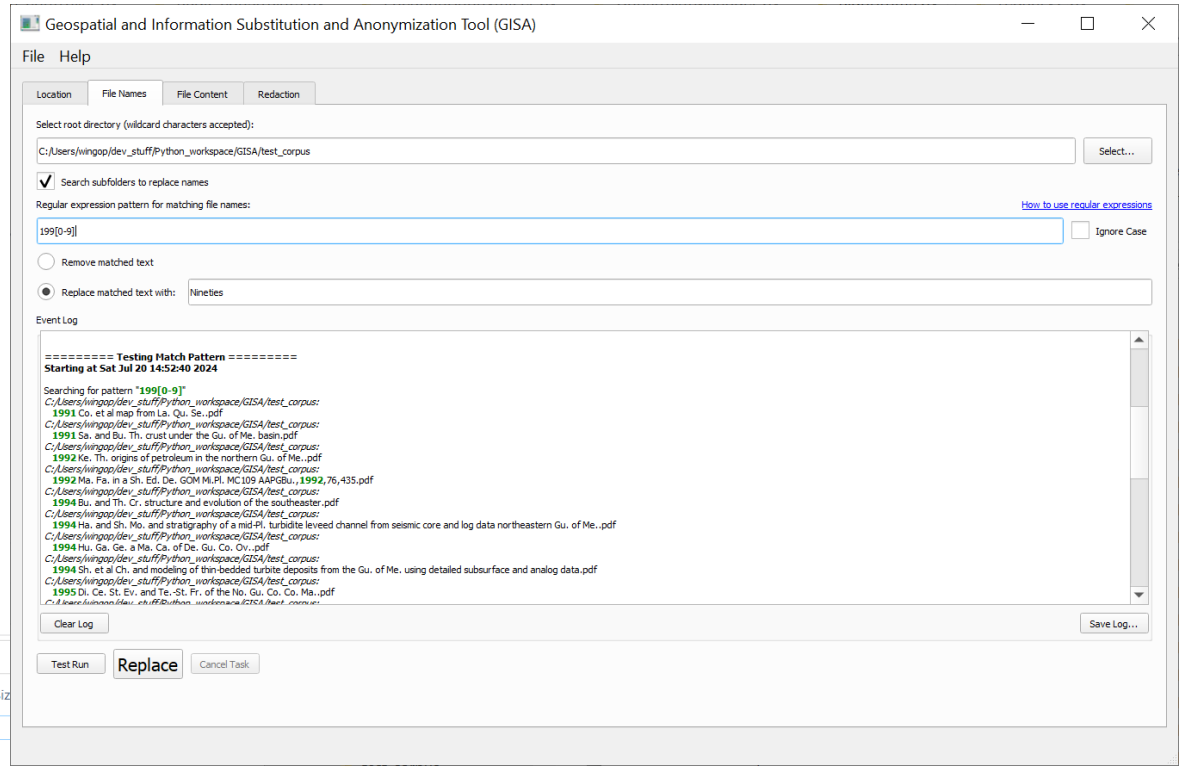
Regular Expression: Syntax used to match patterns and permutations in sequence of values

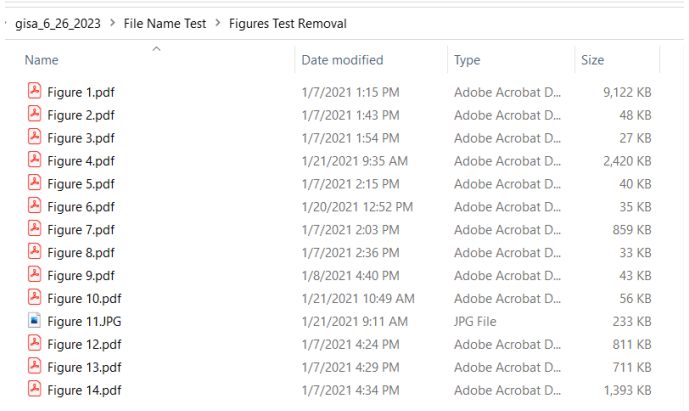
Capture Group: Method to capture portion of RegEx match to be re-inserted in substitution sequence.

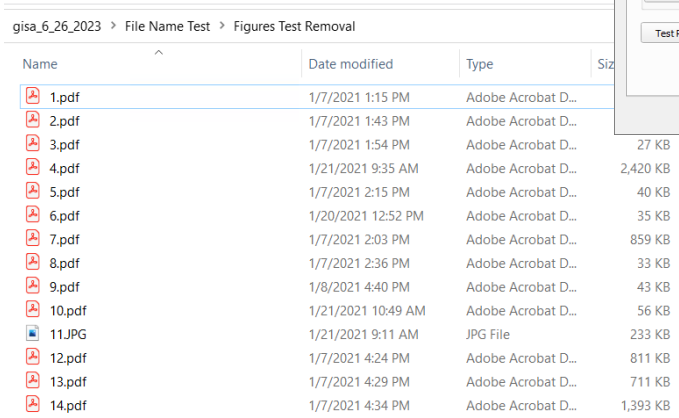
File Name/Content Substitution Anonymization

Batch File renaming

- Select root directory (can use glob statements)
- Supply RegEx pattern for capture
- Two forms of substitution:
  - Remove matched text
  - Replace text (supports RegEx capture groups)





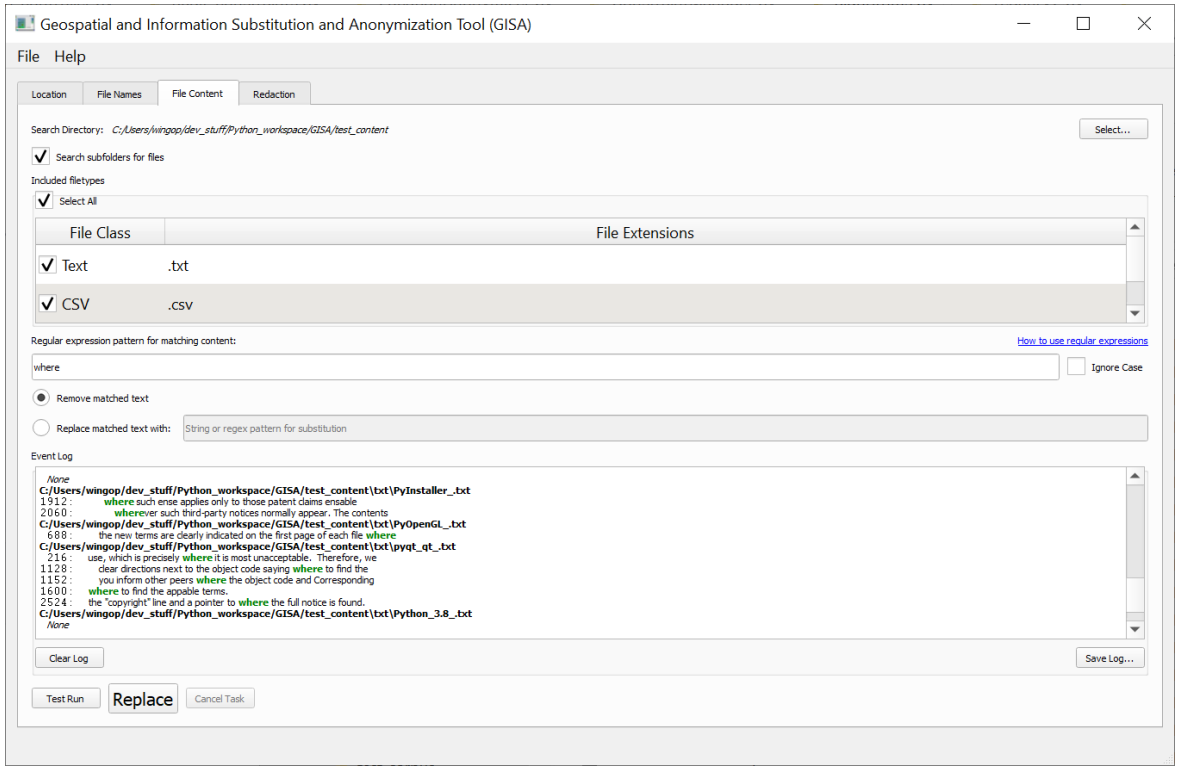


Before and after removal of term “Figure”

File Name/Content Substitution Anonymization

Find/replace in select file types

- Specify root directory.
- Select file types:
  - *Text*
  - *CSV*
  - *Excel**
- Can add custom extensions to handlers
- RegEx pattern for match
- Two forms of substitution:
  - Remove matched text
  - Replace text (supports RegEx capture groups).



Original

Test Batch File 1		
ID	Country	Company
101	United States	Company 1
102	United States	Company 2
103	United States	Company 3
104	United States	Company 4
105	United States	Company 5
106	United States	Company 6
107	United States	Company 7
108	United States	Company 8
109	United States	Company 9
110	United States	Company 10
111	United States	Company 11
112	United States	Company 12
113	United States	Company 13

"United States" Removed

Test Batch File 1		
ID	Country	Company
101		Company 1
102		Company 2
103		Company 3
104		Company 4
105		Company 5
106		Company 6
107		Company 7
108		Company 8
109		Company 9
110		Company 10
111		Company 11
112		Company 12
113		Company 13

"United States" Replaced with "Redacted"

Test Batch File 1		
ID	Country	Company
101	Redacted	Company 1
102	Redacted	Company 2
103	Redacted	Company 3
104	Redacted	Company 4
105	Redacted	Company 5
106	Redacted	Company 6
107	Redacted	Company 7
108	Redacted	Company 8
109	Redacted	Company 9
110	Redacted	Company 10
111	Redacted	Company 11
112	Redacted	Company 12
113	Redacted	Company 13

Redaction Anonymization

Remove Sensitive information without removing its presence.

Redaction is intended for Read-only reports

- Black boxes replace redacted information in copy of document
- The redacted copy does not contain the information specified to be redacted

GISA offers two types of redaction:

- Images: Logos, figures, drawings, etc.
- Text: proper Nouns, adjectives, acronyms, etc.

Redaction Procedure is as follows:

1. User requests GISA to evaluated a PDF document.
2. GISA provides a list of redactable items.
3. User selects specific items to redact.
4. A copy of the PDF produced with the requested data redacted.

Redaction Anonymization

Image Redaction

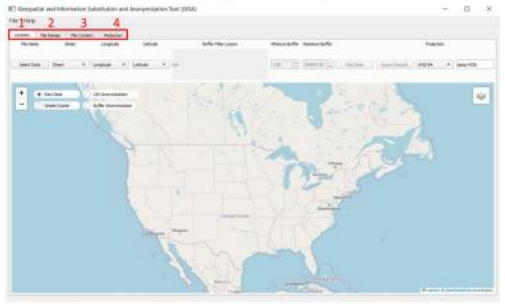
PyMuPDF

https://pymupdf.readthedocs.io/en/latest/

GISA User Interface

Main Window

The following image shows GISA interface. There are four main tabs in GISA.



1. Location – Location tab anonymizes location (latitude/longitude) information.
2. File Names – File names tab anonymizes file names based on user input.
3. File Content – File content tab anonymizes file content tab based on user input.
4. Redaction – Redaction tab performs redaction based on user input.

Prompt Window

A prompt window will open that provides information about the tool while it is running.

****Warning: If you close the prompt window, the GISA tool will close. Leave open while using tool.**

6

U.S. DEPARTMENT OF ENERGY | NATIONAL ENERGY TECHNOLOGY LABORATORY

GISA User Interface

Main Window

The following image shows GISA interface. There are four main tabs in GISA.

1. Location – Location tab anonymizes location (latitude/longitude) information.
2. File Names – File names tab anonymizes file names based on user input.
3. File Content – File content tab anonymizes file content tab based on user input.
4. Redaction – Redaction tab performs redaction based on user input.

Prompt Window

A prompt window will open that provides information about the tool while it is running.

****Warning: If you close the prompt window, the GISA tool will close. Leave open while using tool.**

[redacted]

6

Original PDF → Selected Criteria → Redacted PDF

Redaction Anonymization

Text Redaction

Luke (NLP)
https://github.com/studio-ousia/luke

NHDWaterbody	Surface Waterbodies (lakes, ponds, etc.)	Surface bodies of water (lakes, ponds, etc.) from the National Hydrography Dataset.	shapefile	U.S. Geological Survey, 2019, National Hydrography Dataset (ver. USGS National Hydrography Dataset Best Resolution (NHD) for Hydrologic Unit (HU) 4 - 2001 (published 20191002)), accessed October 10, 2022 at URL https://www.usgs.gov/national-hydrography/access-national-hydrography-products	Do not place within
North_America_Sedimentary_Basins	North America Sedimentary Basins	North America sedimentary basins as defined in the National Carbon Sequestration Atlas V, 2015	shapefile	National Carbon Sequestration Database, 2015, https://edx.netl.doe.gov/dataset/natcarb-alldata-v1502 Bauer, J., Rowan, C., Barkhurst A., Digiulio J., Jones K., Sabbatino M., Rose K., Wingo P. Natcarb, 2018-09-27, https://edx.netl.doe.gov/dataset/natcarb, DOI: 10.18141/1474110	Stay within
Oil_Gas_Atlas5_v3_final	Oil and Gas Field locations	Oil and gas field locations published in the National Carbon Storage Atlas and Database, 2015.	shapefile	National Carbon Sequestration Database, 2015, https://edx.netl.doe.gov/dataset/natcarb-alldata-v1502 Bauer, J., Rowan, C., Barkhurst A., Digiulio J., Jones K., Sabbatino M., Rose K., Wingo P. Natcarb, 2018-09-27, https://edx.netl.doe.gov/dataset/natcarb, DOI: 10.18141/1474110	Stay within
SGMC_Geology	The State Geologic Map Compilation (SGMC) Geodatabase of the Conterminous United States	The State Geologic Map Compilation (SGMC) geodatabase of the conterminous United States (https://doi.org/10.5066/F7WH2N65) represents a seamless, spatial database of 48 State geologic maps that range from 1:50,000 to 1:1,000,000 scale.	shapefile	Horton, J.D., 2017, The State Geologic Map Compilation (SGMC) geodatabase of the conterminous United States (ver. 1.1, August 2017): U.S. Geological Survey data release, https://doi.org/10.5066/F7WH2N65.	Stay within
State	State Boundries	State boundaries for Continental USA.	shapefile	U.S. Census Bureau, Geography Division (2022). US States. Retrieved from https://www.arcgis.com/home/item.html?id=8c2d6d7df8fa4142b0a1211c8dd66903	Stay within
Urban Areas	Urban Areas	Urban Areas	shapefile	U.S. Census Bureau, Geography Division (2021). Urban Areas. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html	Stay within

USGS
edx
geological

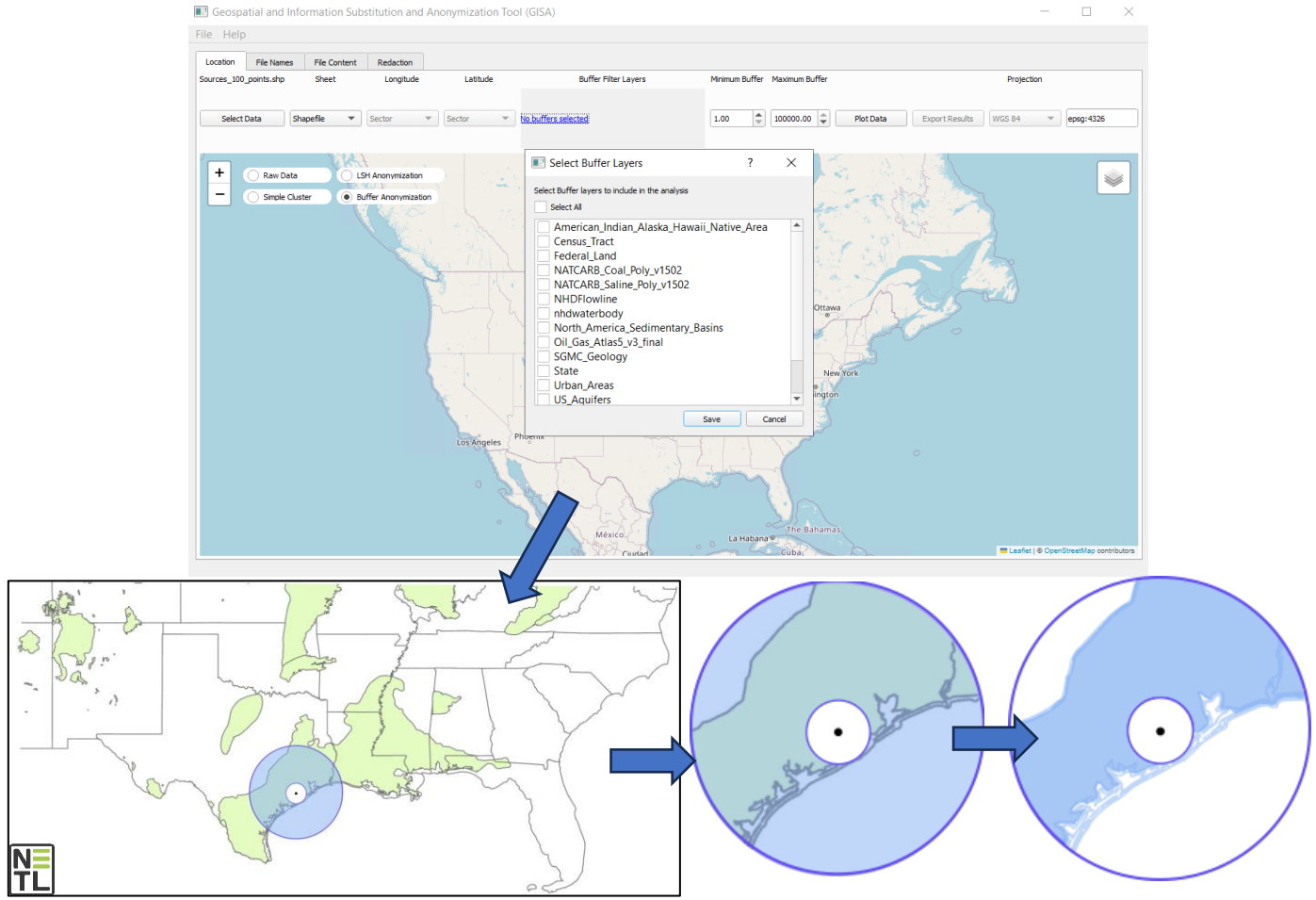
NHDWaterbody	Surface Waterbodies (lakes, ponds, etc.)	Surface bodies of water (lakes, ponds, etc.) from the National Hydrography Dataset.	shapefile	U.S. [redacted] Survey, 2019, National Hydrography Dataset (ver. [redacted] National Hydrography Dataset Best Resolution (NHD) for Hydrologic Unit (HU) 4 - 2001 (published 20191002)), accessed October 10, 2022 at URL [redacted] hydrography/access-national-hydrography-products	Do not place within
North_America_Sedimentary_Basins	North America Sedimentary Basins	North America sedimentary basins as defined in the National Carbon Sequestration Atlas V, 2015	shapefile	National Carbon Sequestration Database, 2015, [redacted] Bauer, J., Rowan, C., Barkhurst A., Digiulio J., Jones K., Sabbatino M., Rose K., Wingo P. Natcarb, 2018-09-27, [redacted] DOI: 10.18141/1474110	Stay within
Oil_Gas_Atlas5_v3_final	Oil and Gas Field locations	Oil and gas field locations published in the National Carbon Storage Atlas and Database, 2015.	shapefile	National Carbon Sequestration Database, 2015, [redacted] Bauer, J., Rowan, C., Barkhurst A., Digiulio J., Jones K., Sabbatino M., Rose K., Wingo P. Natcarb, 2018-09-27, [redacted] DOI: 10.18141/1474110	Stay within
SGMC_Geology	The State Geologic Map Compilation (SGMC) Geodatabase of the Conterminous United States	The State Geologic Map Compilation (SGMC) geodatabase of the conterminous United States (https://doi.org/10.5066/F7WH2N65) represents a seamless, spatial database of 48 State geologic maps that range from 1:50,000 to 1:1,000,000 scale.	shapefile	Horton, J.D., 2017, The State Geologic Map Compilation (SGMC) geodatabase of the conterminous United States (ver. 1.1, August 2017): U.S. [redacted] Survey data release, https://doi.org/10.5066/F7WH2N65.	Stay within
State	State Boundries	State boundaries for Continental USA.	shapefile	U.S. Census Bureau, Geography Division (2022). US States. Retrieved from https://www.arcgis.com/home/item.html?id=8c2d6d7df8fa4142b0a1211c8dd66903	Stay within
Urban Areas	Urban Areas	Urban Areas	shapefile	U.S. Census Bureau, Geography Division (2021). Urban Areas. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html	Stay within

Original PDF → Selected Criteria → Redacted PDF

Lessons Learned

- Data anonymization is multi-faceted
  - Different needs
  - Different types of files
  - Different Relationships

- Anonymization is about trust
  - Industry Partners are more willing to share data if we can demonstrate we can protect it
  - Direct benefit to research space



Next Steps

This project is concluded; however, further goals include:

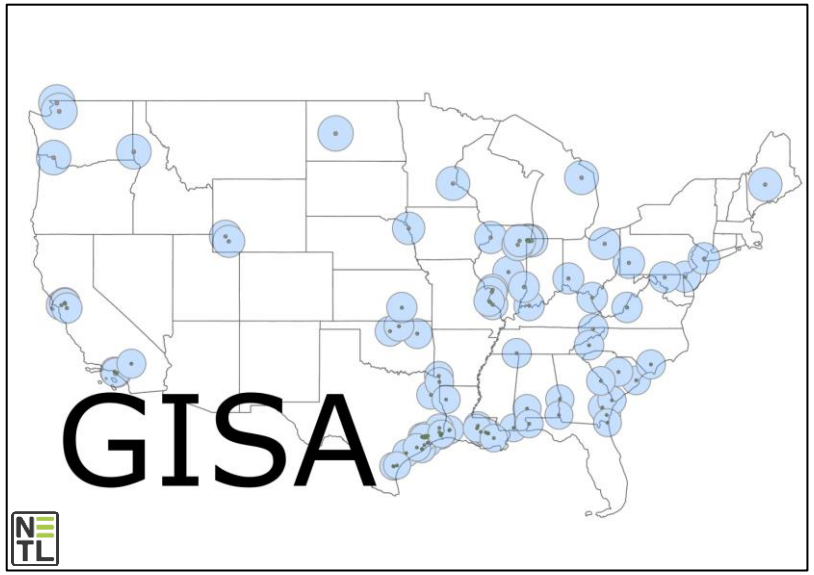
- Proper export of Cluster anonymized geospatial data
- Redaction: ML – aided classification of images
- Further refinement to user interface
- More testing, bug fixes, stability improvements

Acknowledgements

EDX4CCS team

Task 46 team

Battelle Reachback group



Access the GISA tool on EDX!

https://edx.netl.doe.gov/dataset/geospatial-and-information-substitution-and-anonymization-tool-gisa

Citation:
Hoover, B., Gao, M., Wingo, P., Neumann, C., Johnson, C., Lancaster, M., Morkner, P., Sharma, M., Bauer, J., and Rose, K. Geospatial and Information Substitution and Anonymization Tool (GISA) v1.0. National Energy Technology Laboratory, 5/3/2024. www.edx.netl.doe.gov/dataset/geospatial-and-information-substitution-and-anonymization-tool-gisa, DOI: 10.18141/1992880

DEMO & POSTER SESSION

TUESDAY, AUGUST 6, 2024
5:45 PM - 7:45PM
BALLROOM GALLERY

The Geospatial and Information Substitution and Anonymization Tool - GISA

EDX discover
CARBON TRANSPORT & STORAGE DATA AND INNOVATION TO BRIDGE THE DIGITAL DIVIDE

the stats

54 RIC PRESENTATIONS

22 POSTERS

30 TOOL DEMOS

MONDAY

Presentations
(10:30AM - 5:25PM)

- 16 disCO2ver presentations

TUESDAY

Presentations
(10:30AM - 5:45PM)

- 17 SMART presentations
- 2 disCO2ver presentations
- 2 Geographic focus/tool presentations

Posters
(5:45PM - 7:45PM)

- 18 CTS Posters
- 2 PSCC Posters
- 1 CDR Poster
- 1 MLEF Poster

Tool Demos
(5:45PM - 7:45PM)

- 30 Tool Demos
  - SMART
  - NRAP
  - EDX
  - EDX4CCS

WEDNESDAY

Presentations
(2:10PM - 4:30PM)

- 3 transport, research, development, and demonstration activities presentations
- 1 transport modeling presentation
- 1 secure storage (basalts/mafic) presentation

THURSDAY

Presentations
(10:30AM - 5:20PM)

- 8 NRAP presentations
- 2 NETL RIC Presentations
- 2 Offshore presentations

NETL Research · Carbon Storage Program

NRAP National Risk Assessment Partnership

SMART

EDX disco2ver

https://edx.netl.doe.gov/disco2ver

NETL RESOURCES

VISIT US AT: **www.NETL.DOE.gov**

@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory

EDX Support: edxsupport@netl.doe.gov
Patrick Wingo: Patrick.Wingo@netl.doe.gov