Integrating Unstructured Data Analysis into Defense and Intelligence Workflows

James Jones | Scott Cecilio

SEE
WHAT
OTHERS
CAN'T

"Every two days we create as much information as we did up to 2003"

Eric Schmidt, 2010

What does that look like?

Every minute…

Twitter sees new 350,000 tweets

15.2 million Text Messages are sent

600 Wikipedia pages are edited

144 million e-mails are sent

3.6 million Google searches are conducted

Facebook has 510,000 comments posted, 293,000 statuses updated

954,000 new Microsoft Office documents are created

What is Unstructured Data

- Does not have a recognizable structure or is loosely structured
- Can be in a variety of formats and storage mechanisms
  - Word Documents
  - Email
  - Social Media Posts
  - PowerPoint
  - PDF
  - Share drive

Syria Deeply @SyriaDeeply · 7h
The US, UK and France launched airstrikes on multiple sites in #Syria in response to the suspected chemical weapons attack in Eastern Ghouta. It marks the single biggest intervention by western powers against the regime since the start of the conflict.

Syria Deeply @SyriaDeeply · 15h
"Aleppo was built using stone, and that has to continue," @ShaarH @AleppoProject @CEUhungary

Maritime OPINTEL Report.pdf - Adobe Acrobat Reader DC

Home Tools Maritime OPINTEL ... ×

3 / 9 93%

1. (U) CONGO: On 15 January a container ship was boarded at 04:43 S – 011:46 E, at the Pointe Noire Anchorage. Three robbers armed with knives disguised as fishermen in a small boat approached and boarded the anchored container ship. Duty crew spotted the robbers and raised the alarm. On hearing the alarm, the robbers jumped overboard and escaped empty handed in their boat. Port control was contacted, but received no response. (IMB)

2. (U) IVORY COAST: On 23 December the chemical tanker MADONNA 1 was boarded at 05:12 N – 004:02 W, at the Abidjan anchorage. Robbers armed with machine guns and knives boarded an anchored chemical tanker and took the crew hostage. They forced the Chief Engineer

Problems in Integrating Unstructured Data

- Tone can vary wildly
- Not in traditional spatial format
- May or may not contain explicit locational information
- Locational information may take many forms
  - Coordinates
  - Place-names
  - Address

Syria Deeply @SyriaDeeply · 7h

The US, UK and France launched airstrikes on multiple sites in #Syria in response to the suspected chemical weapons attack in Eastern Ghouta. It marks the single biggest intervention by western powers against the regime since the start of the conflict.

Syria Deeply @SyriaDeeply · 15h

"Aleppo was built using stone, and that has to continue," @ShaarH @AleppoProject @CEUhungary

Maritime OPINTEL Report.pdf - Adobe Acrobat Reader DC

Home Tools Maritime OPINTEL ... 3 / 9 93%

1. (U) CONGO: On 15 January a container ship was boarded at 04:43 S – 011:46 E, at the Pointe Noire Anchorage. Three robbers armed with knives disguised as fishermen in a small boat approached and boarded the anchored container ship. Duty crew spotted the robbers and raised the alarm. On hearing the alarm, the robbers jumped overboard and escaped empty handed in their boat. Port control was contacted, but received no response. (IMB)

2. (U) IVORY COAST: On 23 December the chemical tanker MADONNA 1 was boarded at 05:12 N – 004:02 W, at the Abidjan anchorage. Robbers armed with machine guns and knives boarded an anchored chemical tanker and took the crew hostage. They forced the Chief Engineer

How to Integrate Unstructured Data into ArcGIS

Native Esri Capability

Third Party Integration

Coordinates
Custom Locations
User defined keywords

What are you looking for?

Locations
People/Organizations
Events
Dates
Relationships

ArcGIS Pro

What is the best tool?

Natural Language Processing

- Data is at least somewhat understood
- Data benefits from identifiable and repeating patterns
- Little to no programming experience available/needed

How is it best used?

- Data is not well understood
- Data does not contain identifiable and/or repeating patterns
- Integration needed

ArcGIS LocateXT

Extract Locations from Unstructured Data

SEE
WHAT
OTHERS
CAN'T

Extracting Locations with ArcGIS

- LocateXT Extension for ArcGIS Desktop and Enterprise
- Available for ArcMap 9.1 and later
- Available in ArcGIS Pro at 2.3
- 100% Feature function as ArcGIS Pro 2.4
- Uses pattern matching regular expressions (REGEX) to search for coordinates in a variety of formats
- Uses custom location list to match/extract other patterns (place names, codes, other terms)

2.18 Erha Oil Terminal (5°21'N., 4°20'E.), a deep-water facility, lies in the vicinity of the Erha Oil Field and is situated within an exclusion zone bordered by a line joining the following positions:
a. 5°21.7'N, 4°17.7'E.
b. 5°23.9'N, 4°20.4'E.
c. 5°23.9'N, 4°22.5'E.
d. 5°21.7'N, 4°23.6'E.
e. 5°18.0'N, 4°21.5'E.
f. 5°19.0'N, 4°17.7'E.
2.18 The terminal consists of a Floating Production Storage and Off loading (FPSO) vessel and an SBM. The SBM is moored 1 mile SE of the FPSO and is connected to it by two steel pipelines

At 1547Z, one adult female and two children ran from one house IVO 38PMB1143097553 to the pickup and then back into the same house.

At 1732Z, two adult males with long guns ran from the underpass into thick brush north of the underpass IVO 38PMB1091196278.

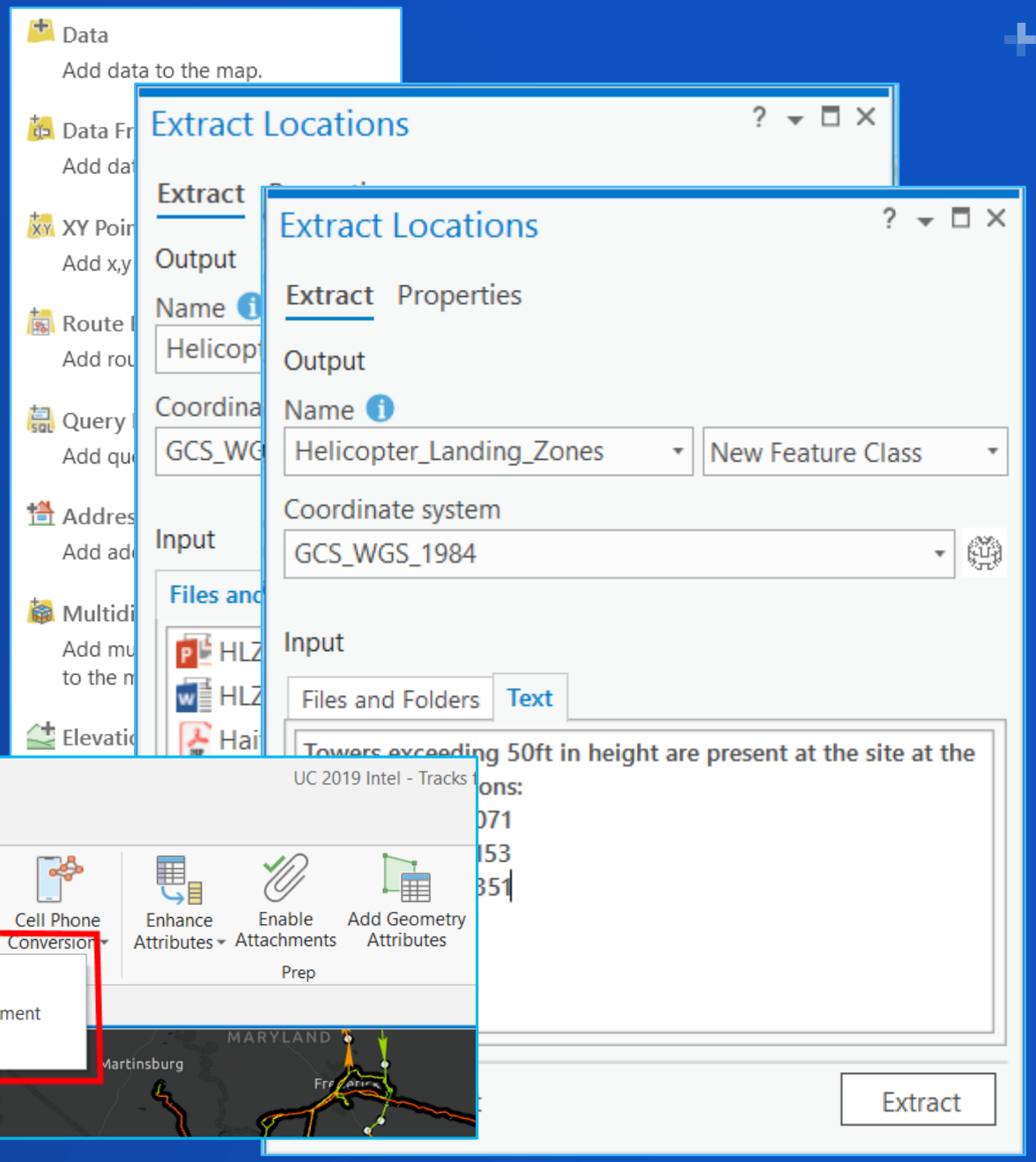
At 1849Z, three adult males were seen covering object with debris on side of road on Highway 101 38PMB1230896427.

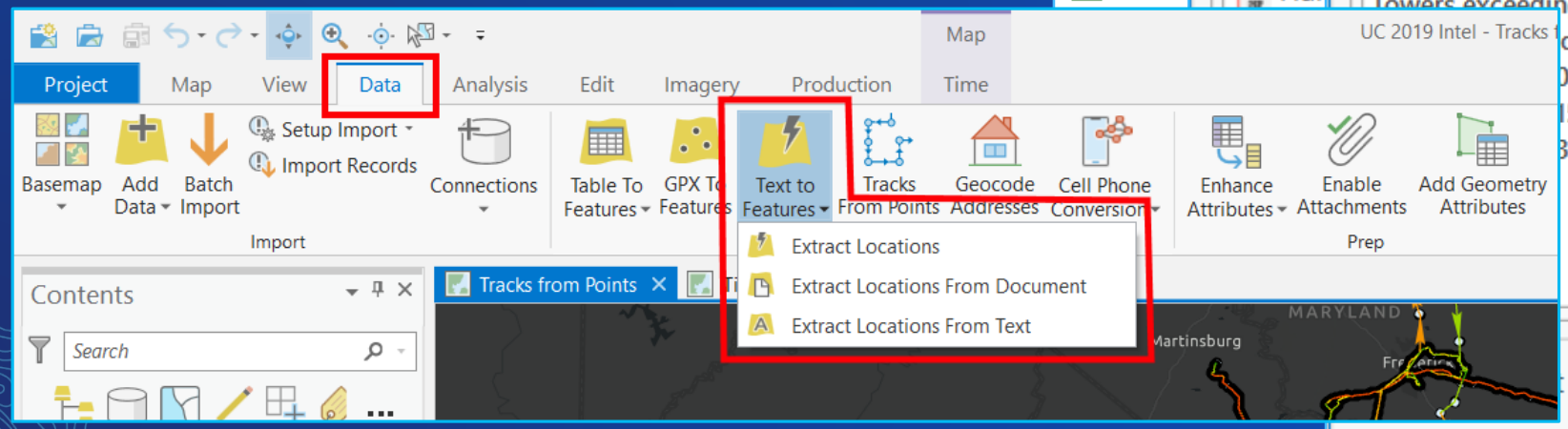
DEPARTURE: **WEDNESDAY 12 SEP** Please verify flight times prior to departure

UNITED AIRLINES UA 5366	ONT ONTARIO, CA ▸ SFO SAN FRANCISCO, CA		Aircraft: CRJ-CANADAIR REGIONAL JET
Operated by: /SKYWEST DBA UNITED EXPRESS	Departing At: 19:35	Arriving At: 21:02	Distance (in Miles): 363
Duration: 1hr(s) 27min(s)	Terminal: TERMINAL 2	Terminal: TERMINAL 3	Stop(s): 0

Extracting Locations in ArcGIS Pro

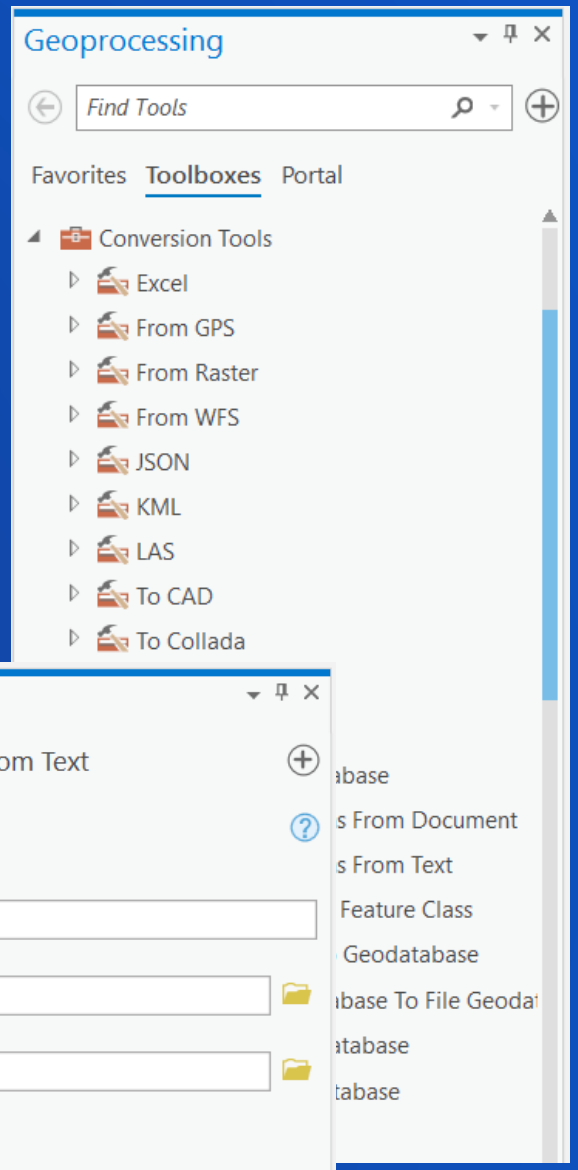
- New option added to the "Add Data" button
- Allows for a user to drag and drop documents or copied text into a window
- Can create a new feature class or append it to an existing one
- Included with ArcGIS Pro for Intelligence

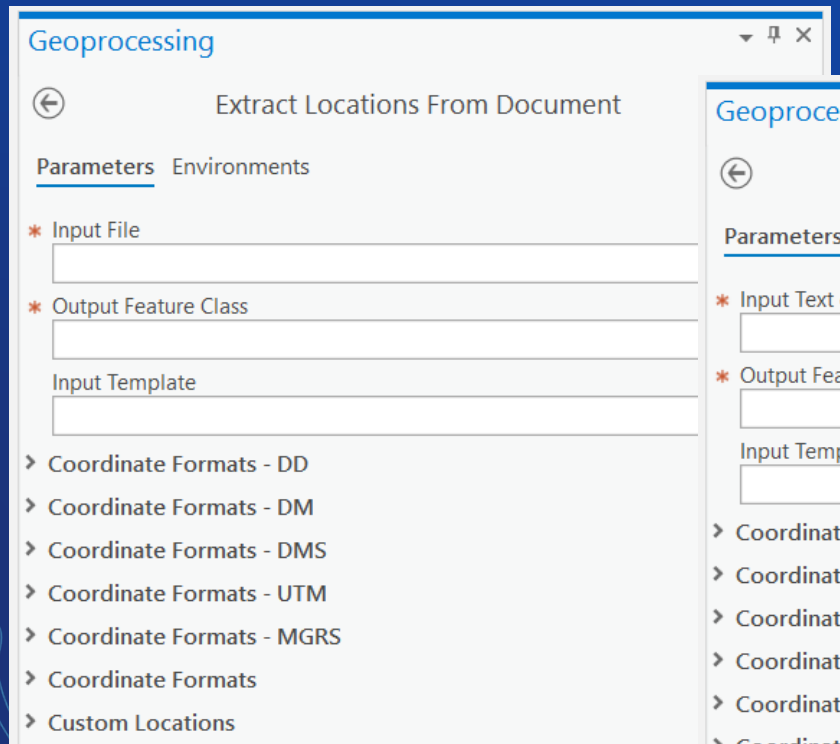


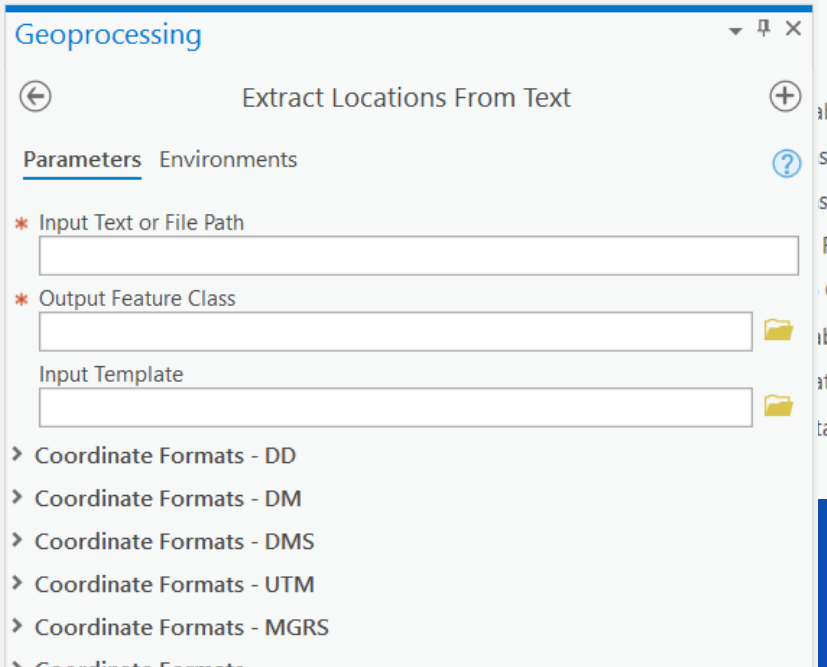


Extracting Locations in ArcGIS Pro

- Two Geoprocessing Tools added
- Located in the ***Conversion Tools*** –> ***To Geodatabase***
  - Extract Locations from Document
  - Extract Locations from Text

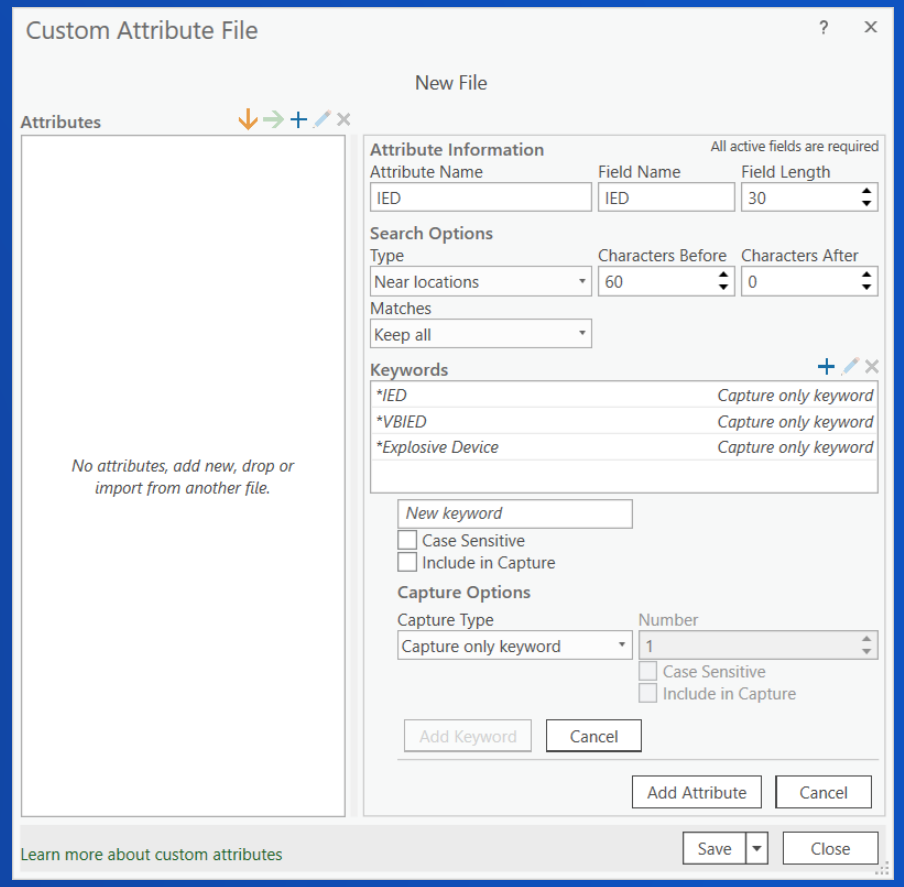






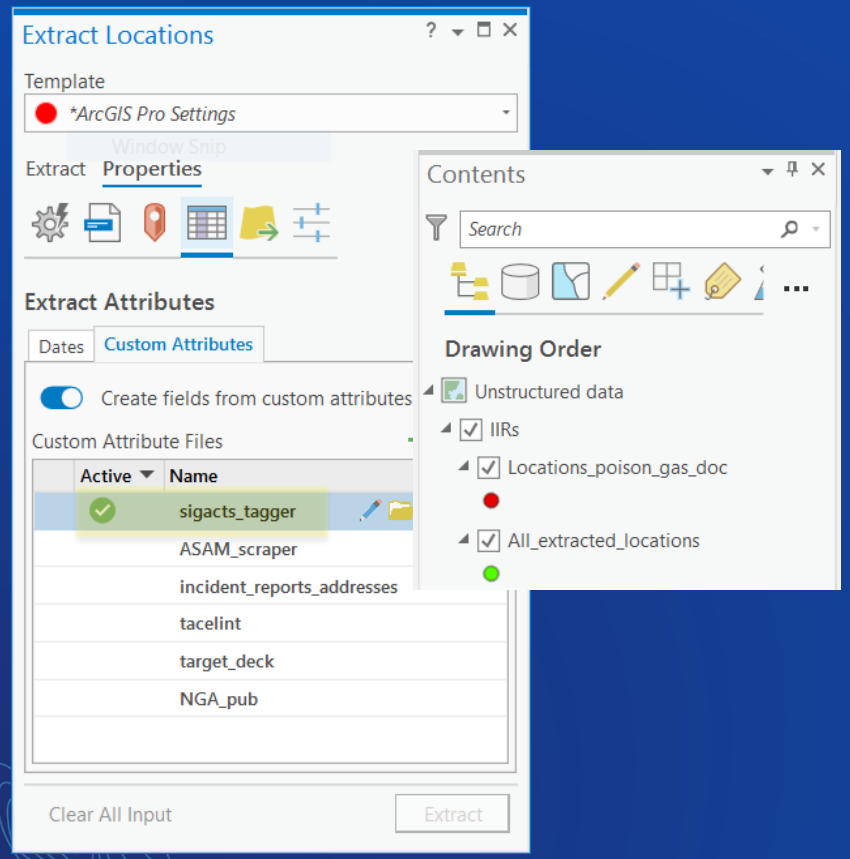
Extracting Custom Attributes

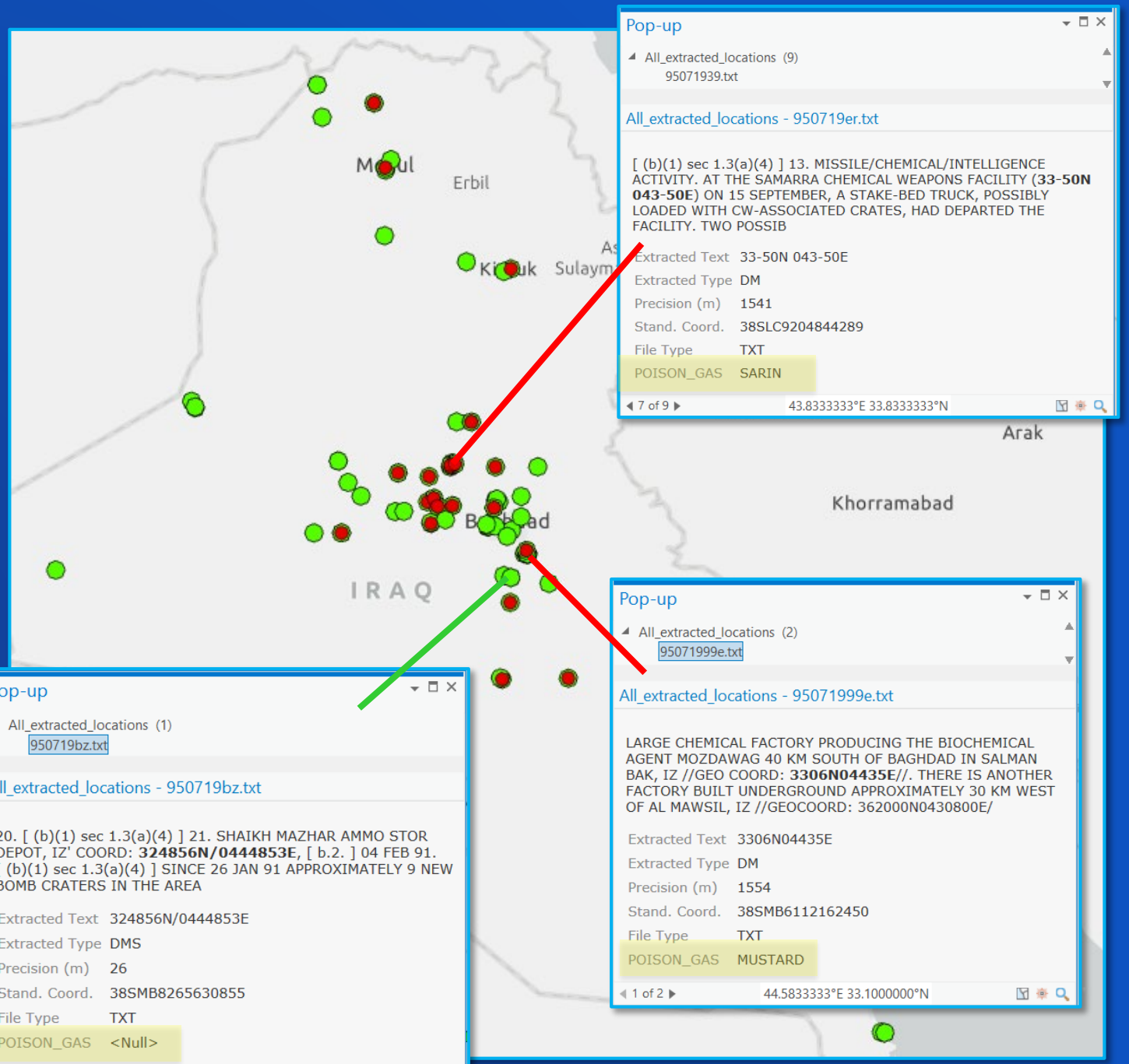
- Ability to create custom attributes based on content within document or near a location
  - Triggered by location extraction
- Based on keywords
  - Tag locations based on keywords
  - Scrape/harvest portions of document based on keywords
- Ability to extract based off of:
  - Number of characters/words
  - Number of lines/blank line
  - Stop string
- Previously built in separate LocateXT desktop application (until Pro 2.4)



Extracting Custom Attributes

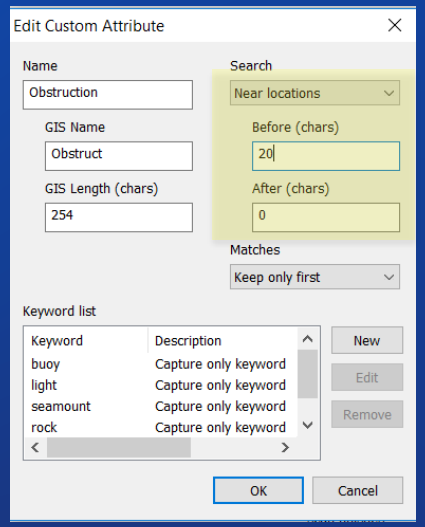
Tag extracted locations based on keyword found in source document

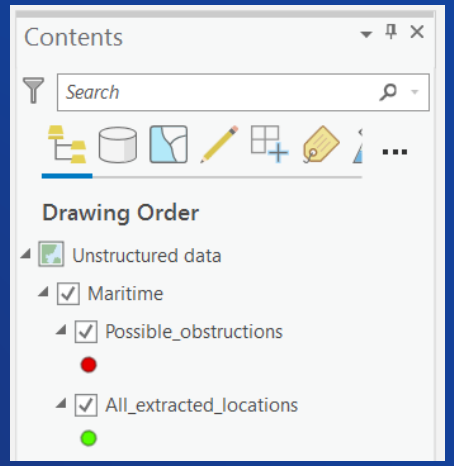


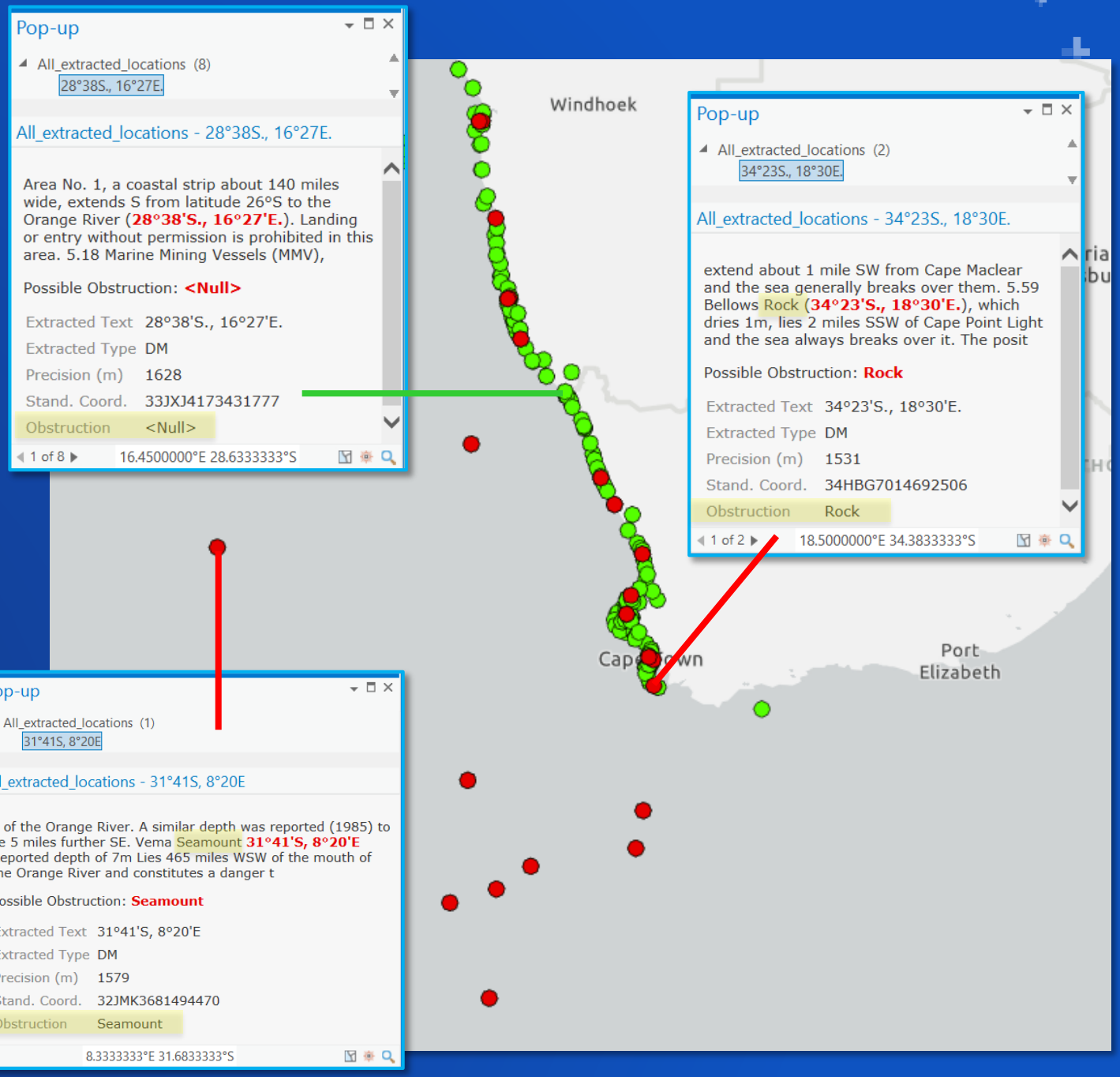


Extracting Custom Attributes

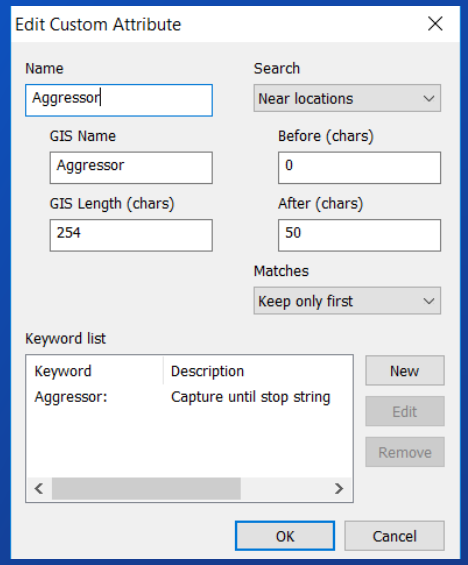
Tag extracted locations based on keyword found in proximity to location

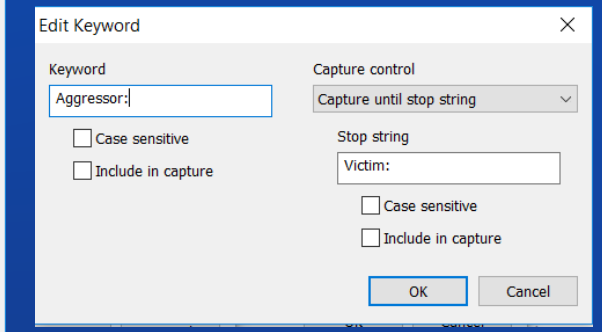


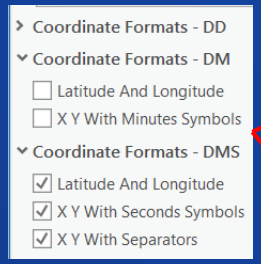


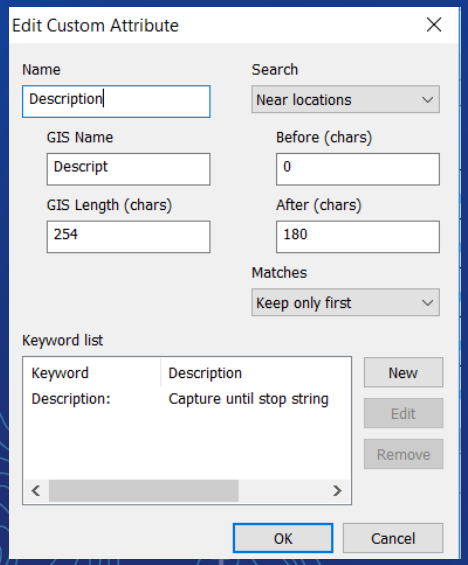


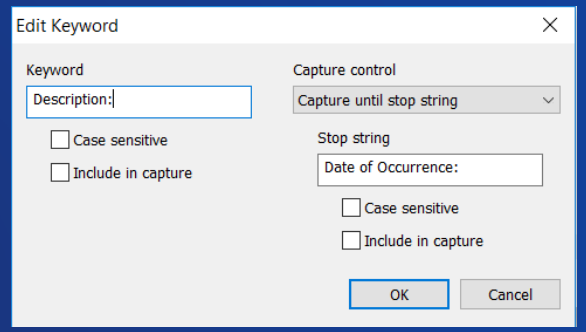
Custom capture text based on keywords found in proximity to location

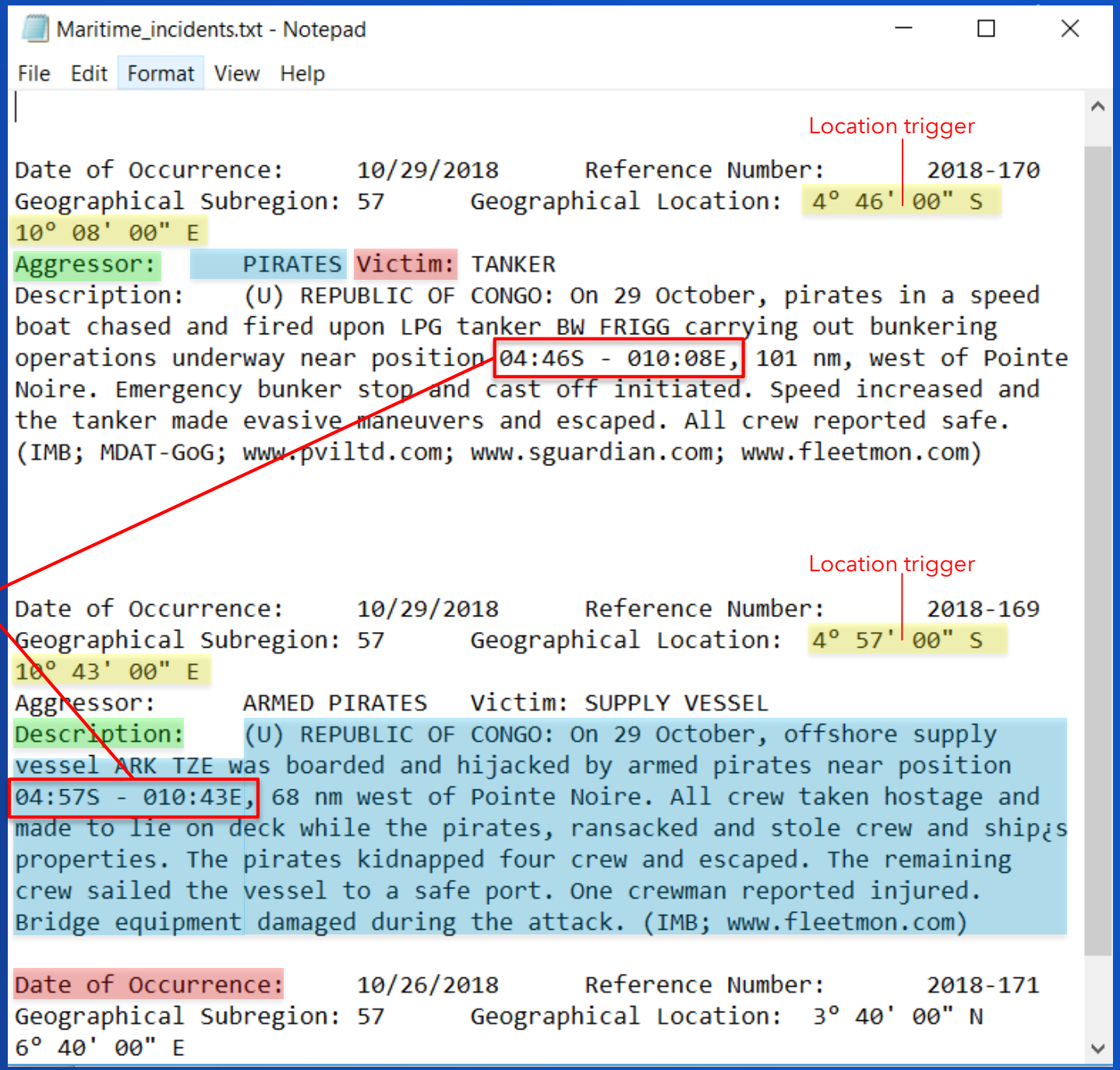




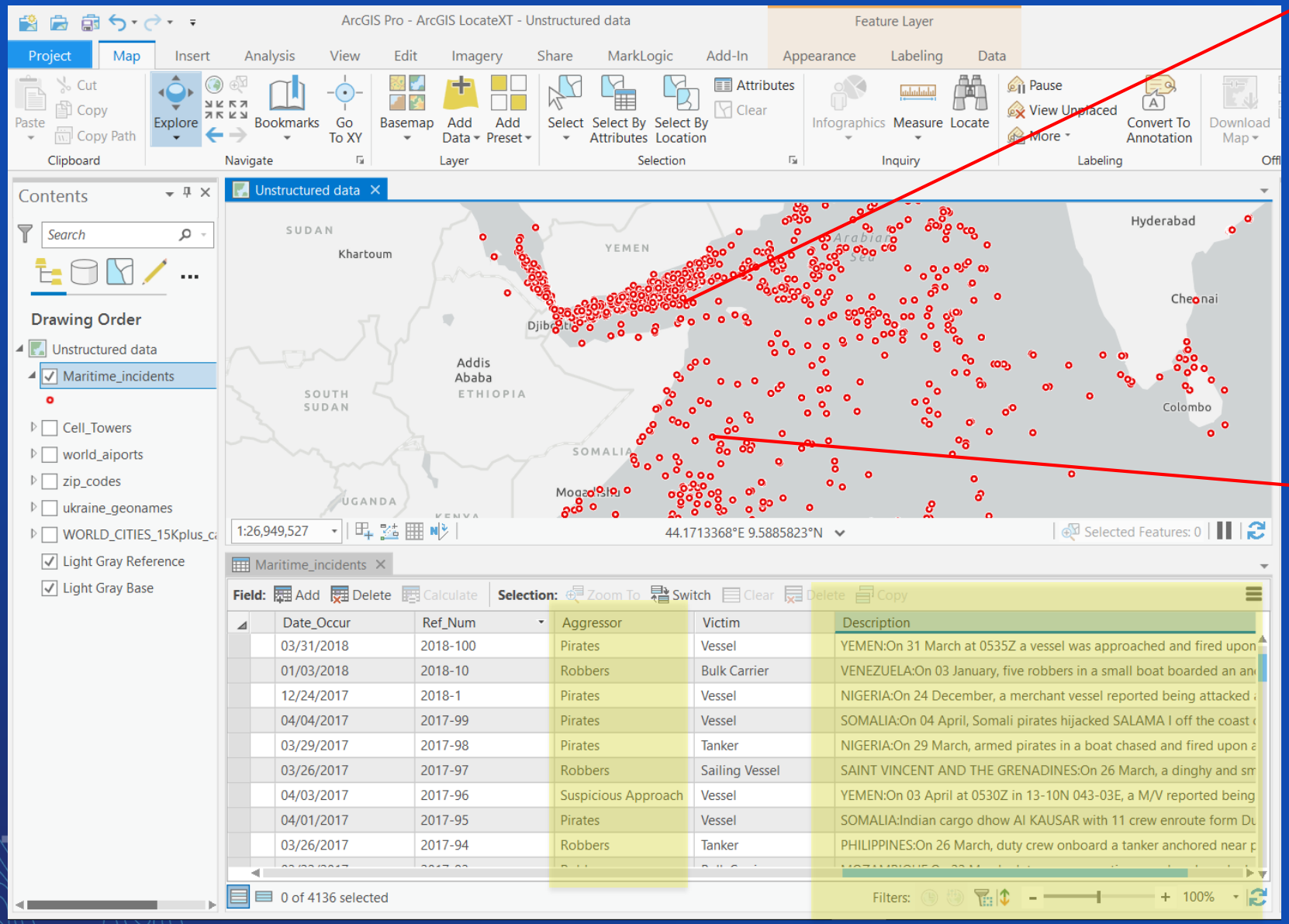


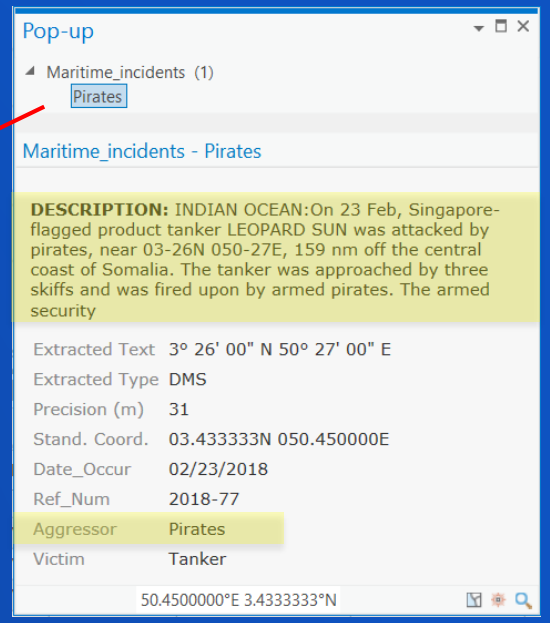


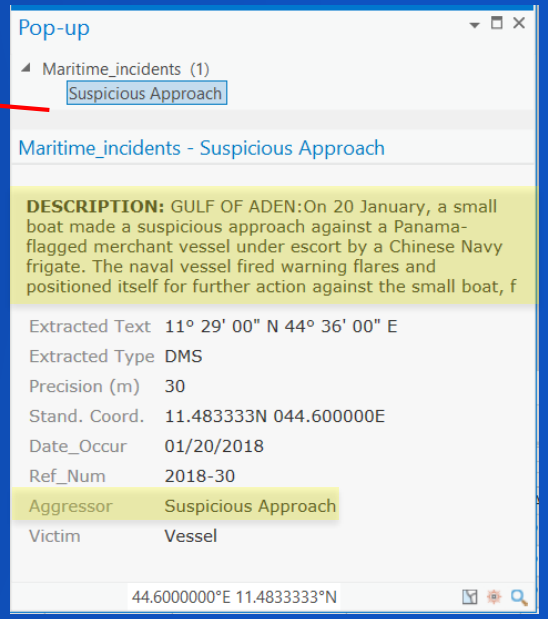


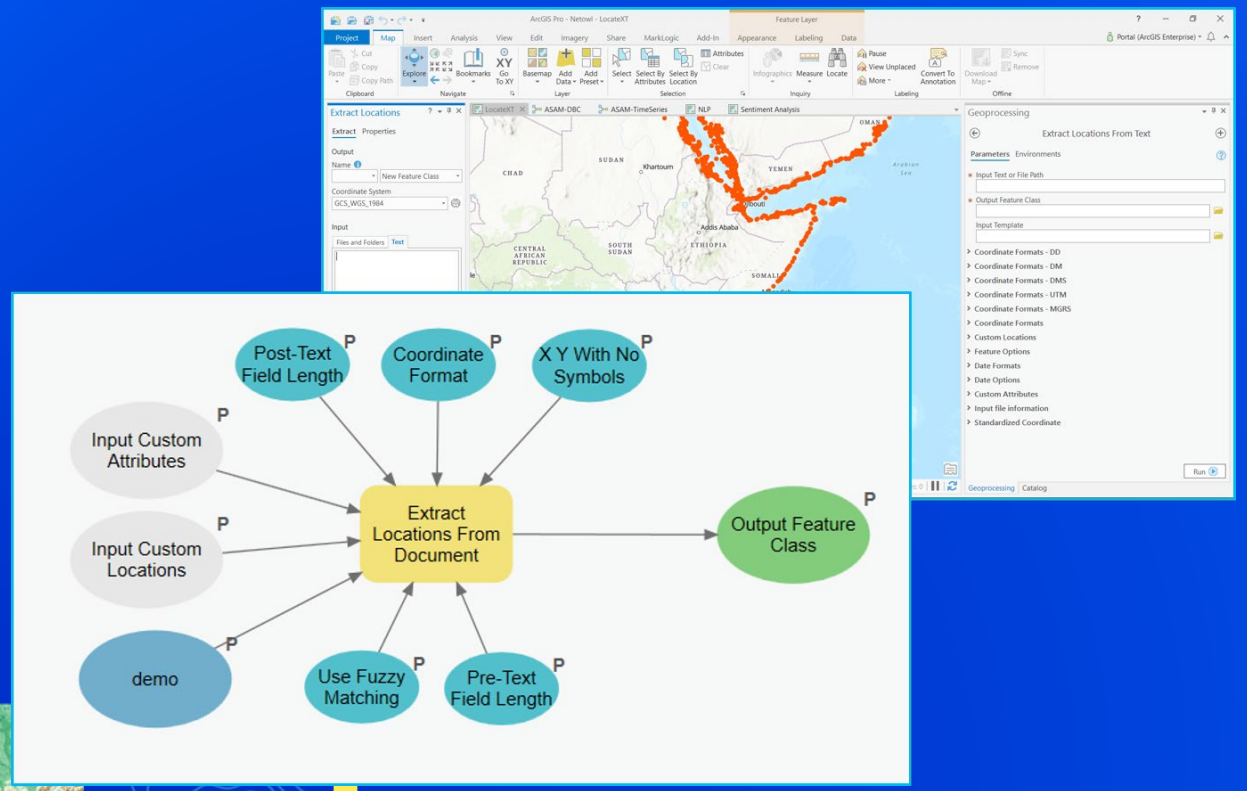


Custom capture text based on keywords found in proximity to location









Explore Unstructured Data through LocateXT and Custom Attributes

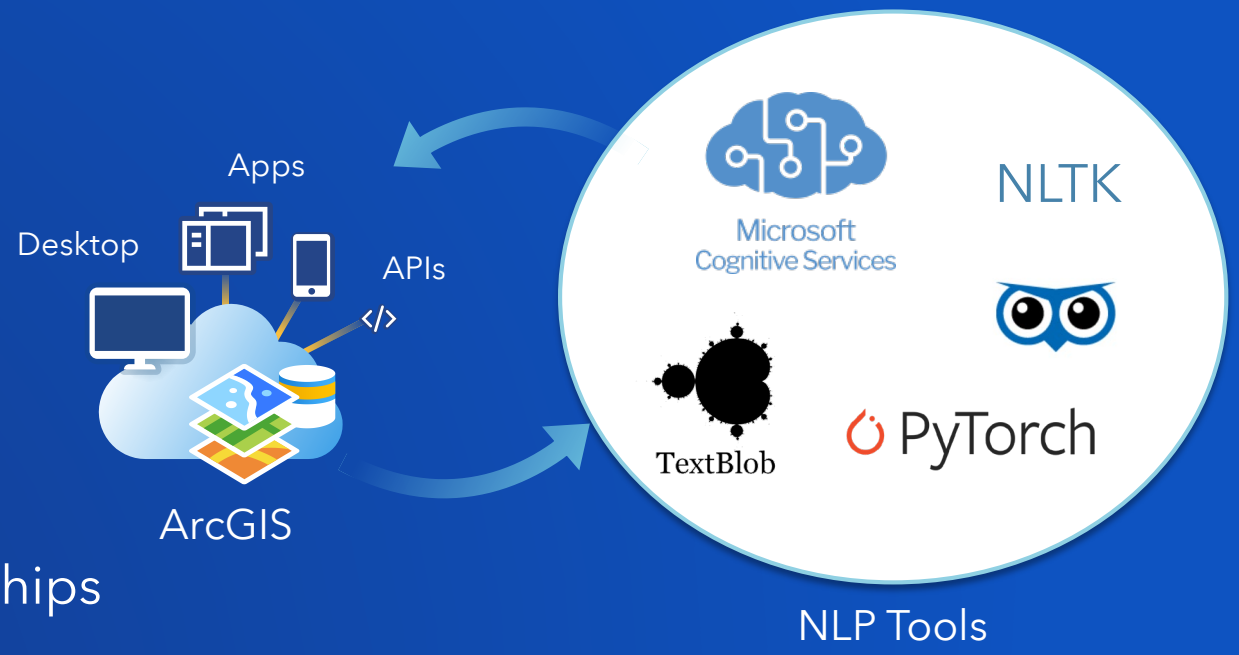
Integrating NLP

with ArcGIS

SEE
WHAT
OTHERS
CAN'T

Background

- What is NLP?
- Main fields of NLP
- Types of NLP solutions
  - Open Source
  - Proprietary
  - as a Service
- Link entities, events, and create relationships
- Organize data into an ontology



Entities and Relationships

bin laden.txt - Notepad

File Edit Format View Help

Osama bin Laden is a terrorist. He is responsible for hundreds of attacks worldwide. Osama bin Laden was born in Saudi Arabia. He bombed the US Embassy in Kenya. Osama bin Laden attacked the World Trade Center at 285 Fulton St, New York, NY 10007. Osama bin Laden was killed in a raid by US Special Forces at a compound location 34 10 9.51N 73 14 32.78E.

Bin Laden a founder of al-Qaeda, the organization responsible for the September 11 attacks in the United States and many other mass-casualty attacks worldwide. Bin Laden's father was Mohammed bin Awad bin Laden, a Saudi millionaire from Hadhramaut, Yemen.

President Clinton targeted bin Laden. On August 20, 1998, 66 cruise missiles launched by United States Navy ships in the Arabian Sea struck bin Laden's training camps near Khost in Afghanistan, missing him by a few hours.

Mohammed Atta, and associate of bin Laden and also part of the al-Qaeda terrorist organization, live in Miami for a while at a residence 6010 SW 8th St, West Miami, FL 33144.

Another Bin Laden associate, Abu Musab al-Zarqawi, was killed in a targeted killing on June 7, 2006, while attending a meeting in an isolated safehouse approximately 5 miles northwest of Baqubah. At 14:15 GMT, two United States Air Force F-16C jets identified the house and the lead jet dropped two 500-pound (230 kg) guided bombs, a laser-guided GBU-12 and GPS-guided GBU-38 on the building located at 33 48 02.83N 44 30 48.58E. Five others were also reported killed.

Entities (spatial)

- Saudi Arabia
- 285 Fulton St, New York,NY 10007
- 34 10 9.51N 73 14 32.78E
- Hadhramaut, Yemen
- approximately 5 miles northwest of Baqubah

Entities (non-spatial)

- Osama bin Laden
- Terrorist
- US Embassy
- US Special Forces
- August 20, 1998
- 66 cruise missiles

Links

- Osama Bin laden -- Saudi Arabia (birthplace)
- US Embassy -- Kenya

Events

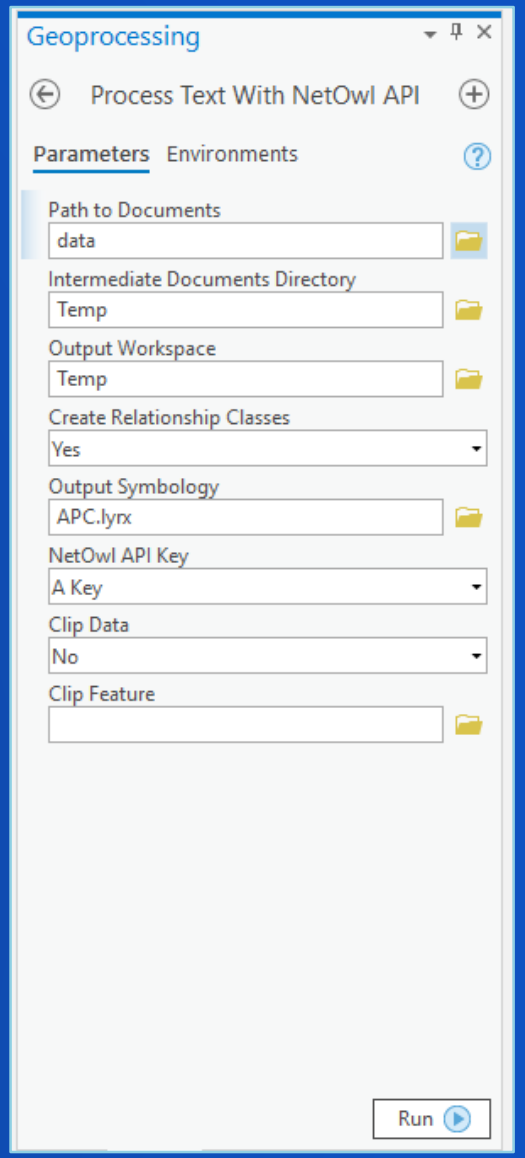
- Osama bin Laden attacked World Trade Center
- Abu Musab al-Zarqawi was killed June 7, 2006

Possible Use Cases of Unstructured Data

- Deriving locations from text
  - Named Entity Recognition
- Analyzing and enhancing existing spatial data containing attributes with free-text narrative
  - Semantic relationship extraction
  - Sentiment Classification (Deep Learning and Bayesian)
  - Topic Identification
  - Machine Translation (Deep Learning and Traditional)

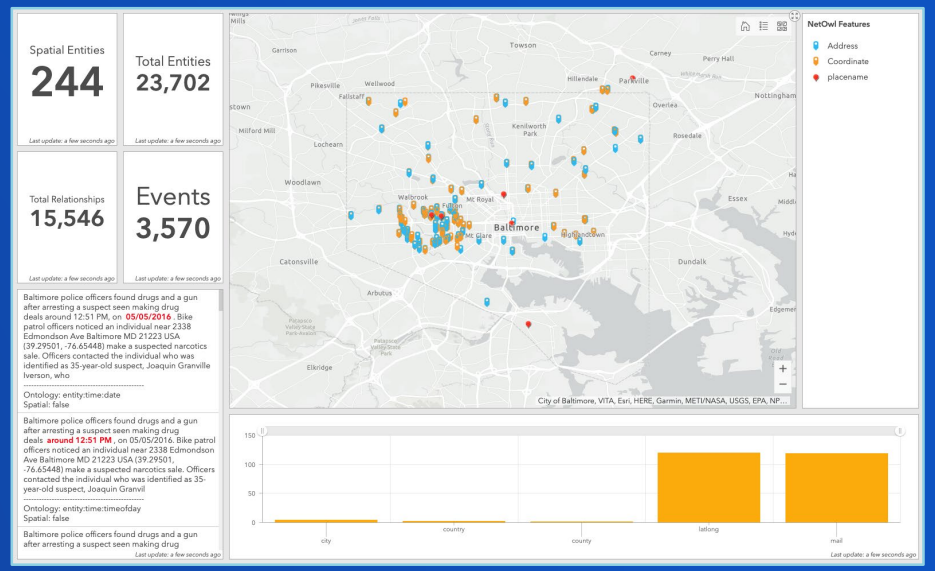
Integrating NLP Capabilities with ArcGIS Pro

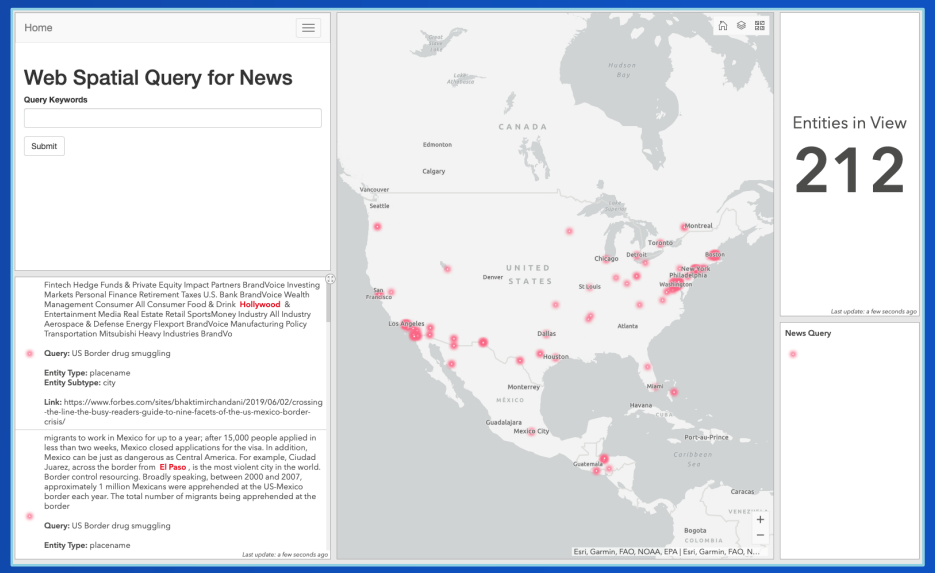
- Many NLP offerings have robust support for Python
- Integrates near seamlessly with ArcPy
- Create Python Toolboxes/Script Tools
- Allows to extract relevant data based on data local to their machine
- Integrates seamlessly into users existing workflows



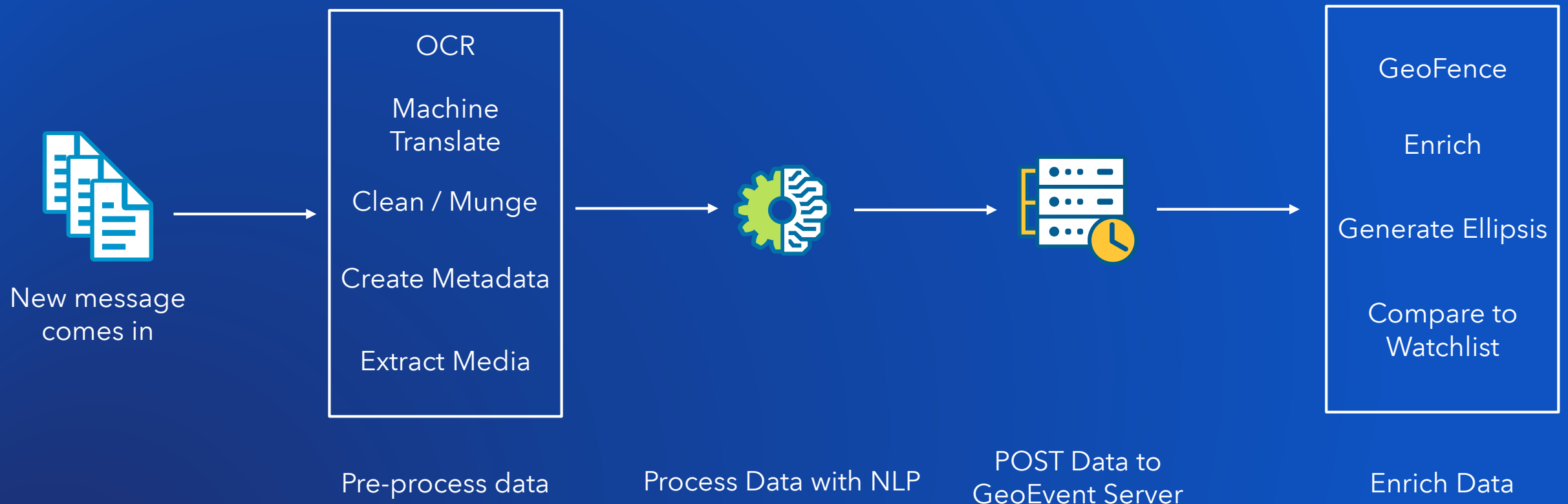
Integrating NLP Capabilities with ArcGIS Enterprise

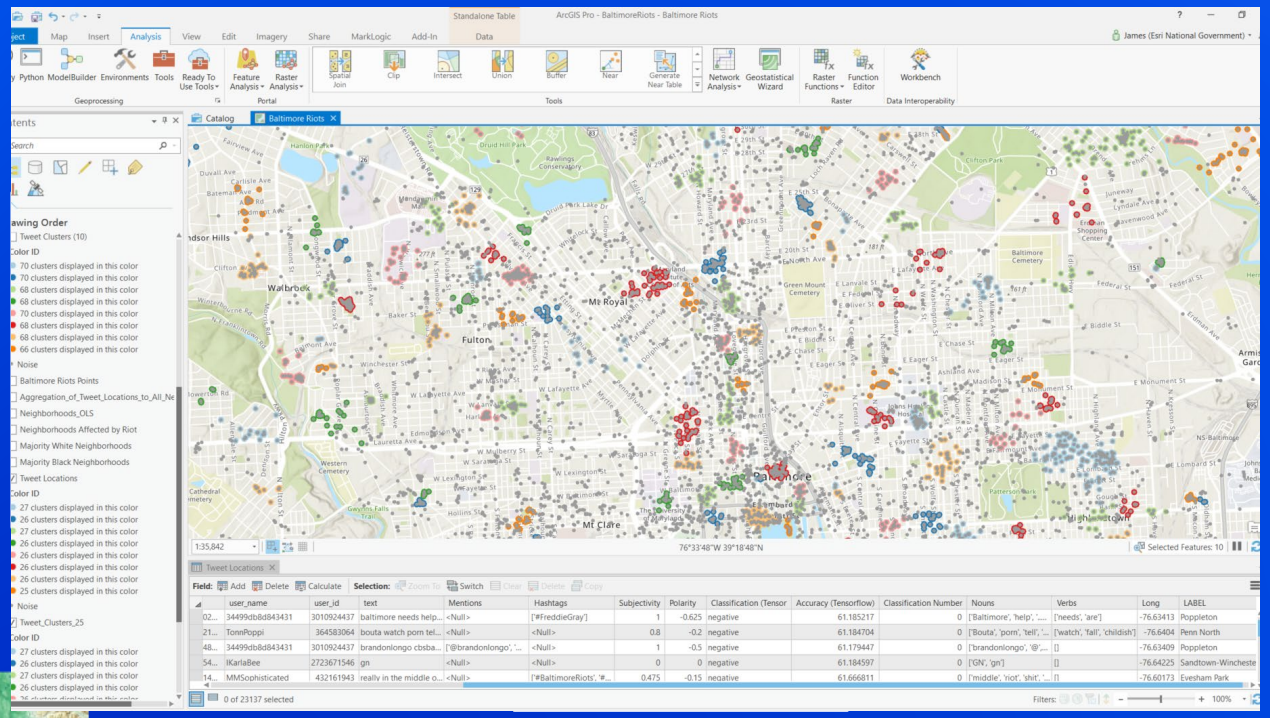
- Process data once, enable users at all levels access to data*
- Integrate data into an enterprise geodatabase, persist entity links using relationship classes
- Construct a robust unstructured data pipeline
  - Allows for tailored extraction based on organizational needs
  - Simplifies integration with other enterprise programs
  - Integration of other components (Machine Translation, OCR, etc.)





Building an Unstructured Data Pipeline

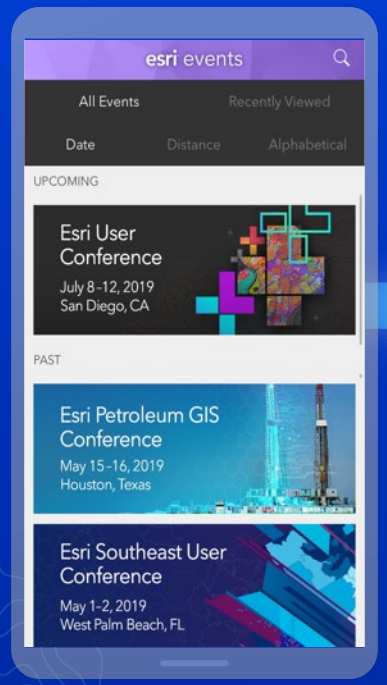




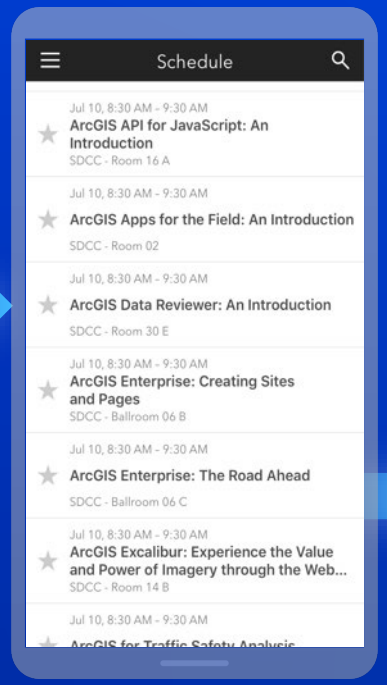
NLP Integration with ArcGIS

Please Share Your Feedback in the App

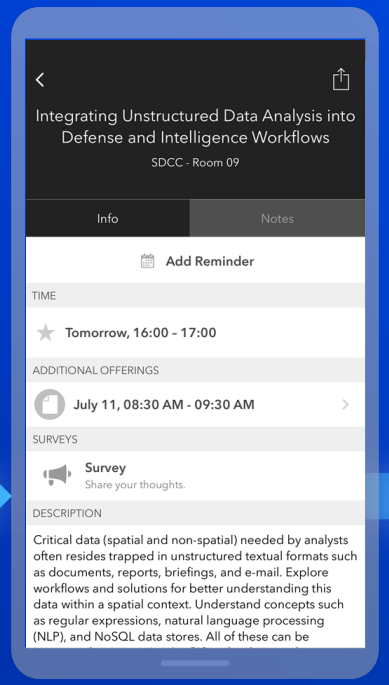
Download the Esri Events app and find your event



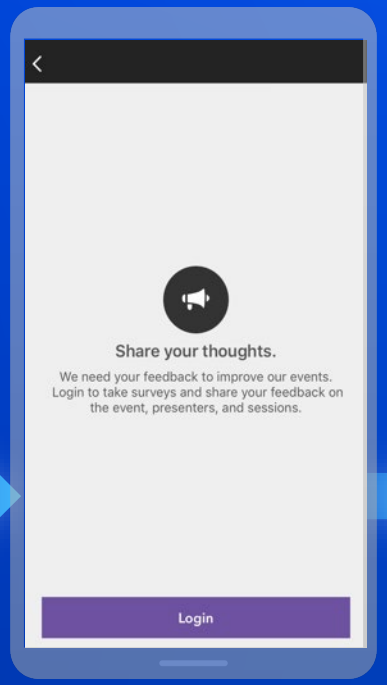
Select the session you attended



Scroll down to "Survey"



Log in to access the survey



Complete the survey and select "Submit"