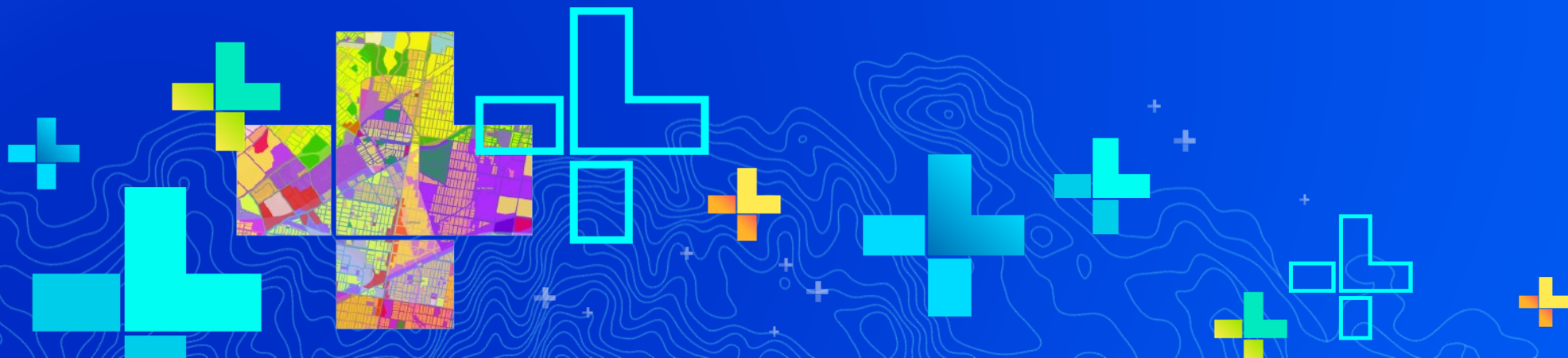





Integrating Unstructured Data Analysis into Defense and Intelligence Workflows

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

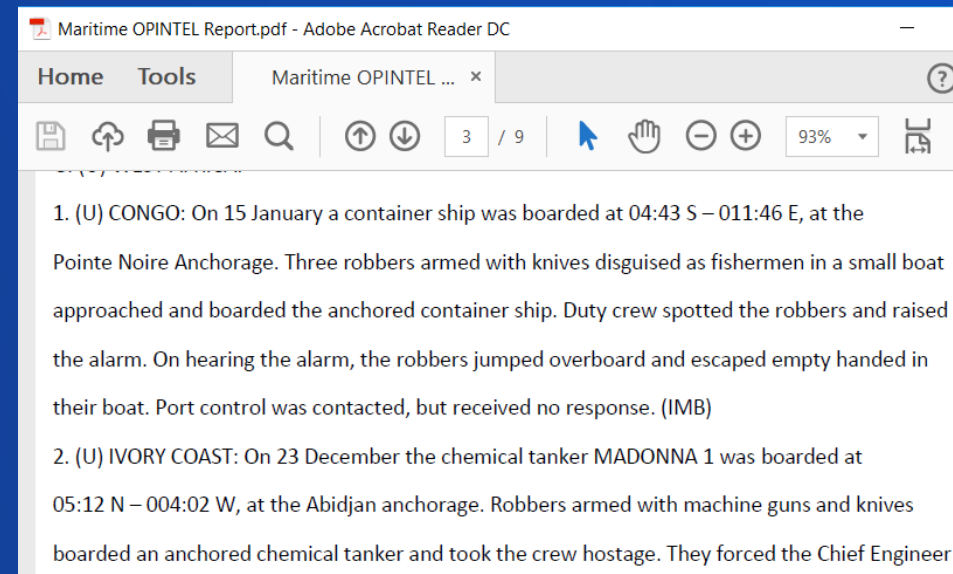
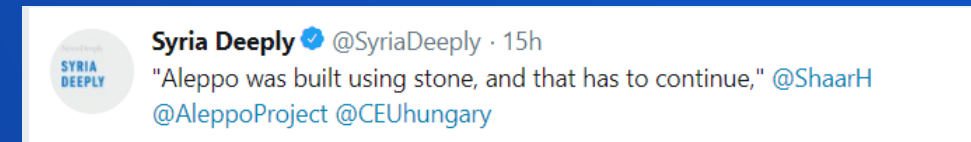
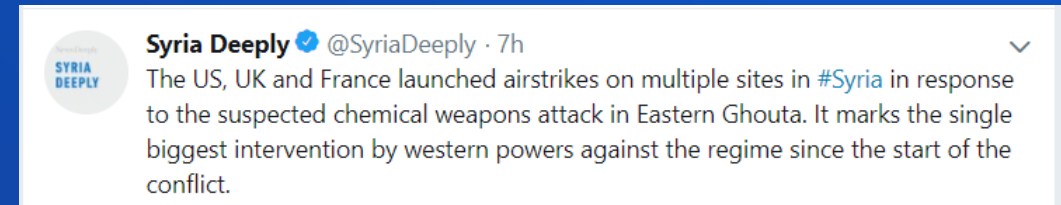


954,000 new Microsoft Office documents
are created



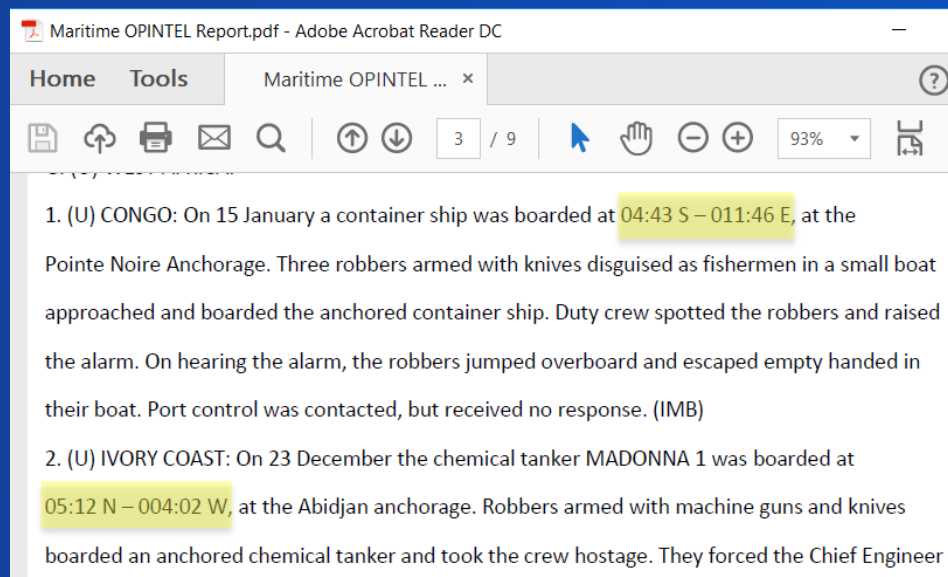
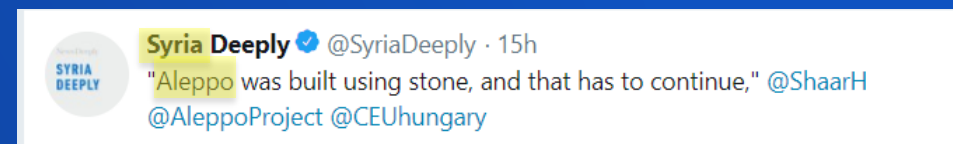
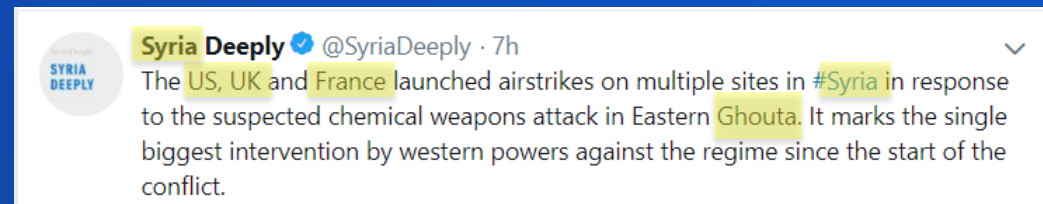
What is Unstructured Data

- 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

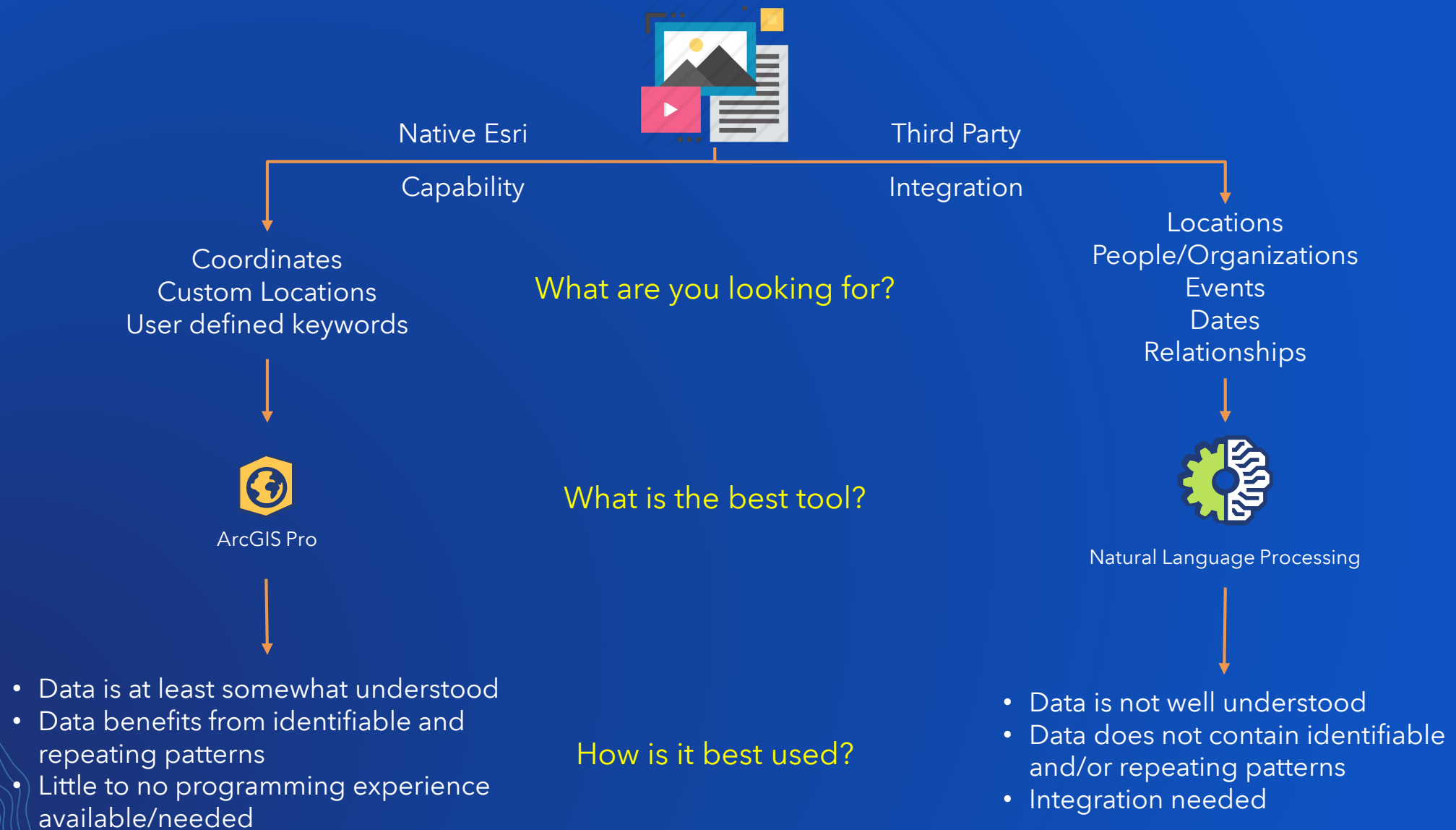


Problems in Integrating Unstructured Data

- 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



How to Integrate Unstructured Data into ArcGIS





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:

- 5°21.7'N, 4°17.7'E.
- 5°23.9'N, 4°20.4'E.
- 5°23.9'N, 4°22.5'E.
- 5°21.7'N, 4°23.6'E.
- 5°18.0'N, 4°21.5'E.
- 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

Operated by:
/SKYWEST DBA UNITED
EXPRESS

Duration:
1hr(s) 27min(s)

ONT
ONTARIO, CA

SFO
SAN FRANCISCO, CA

Departing At:
19:35

Terminal:
TERMINAL 2

Arriving At:
21:02

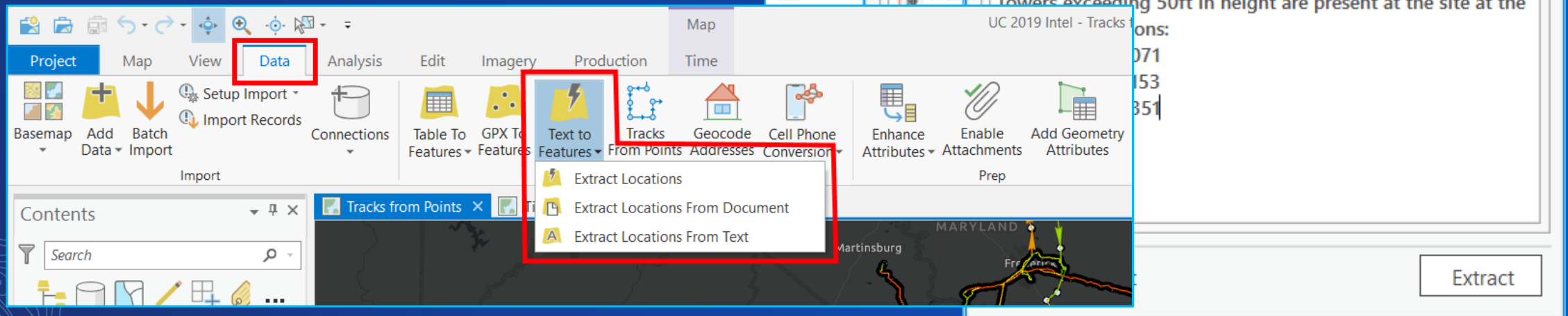
Terminal:
TERMINAL 3

Aircraft:
CRJ-CANADAIR
REGIONAL JET

Distance (in Miles): 363
Stop(s): 0

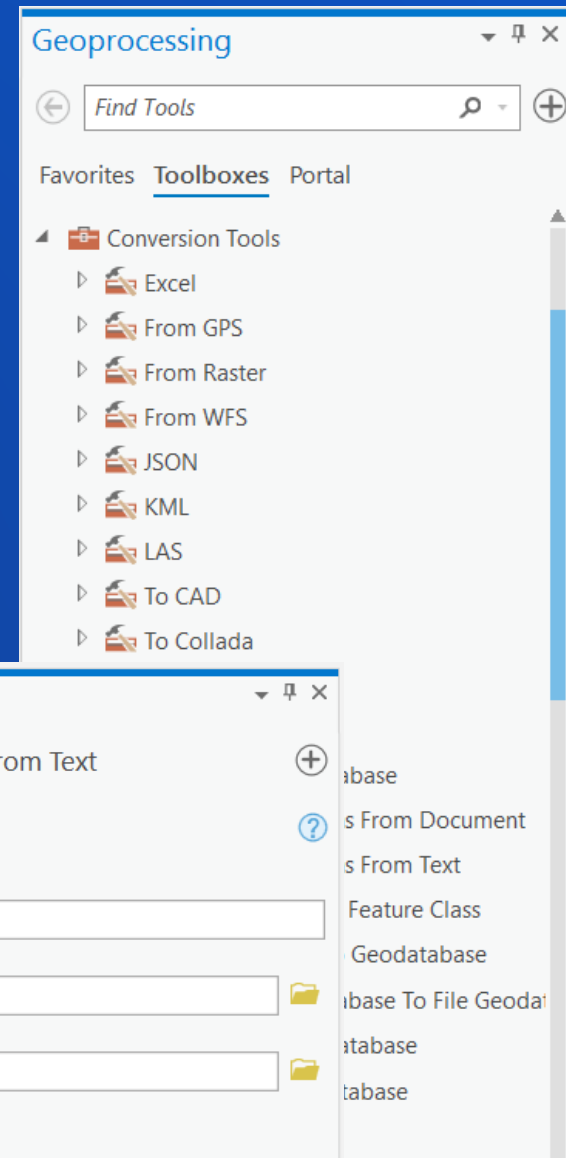
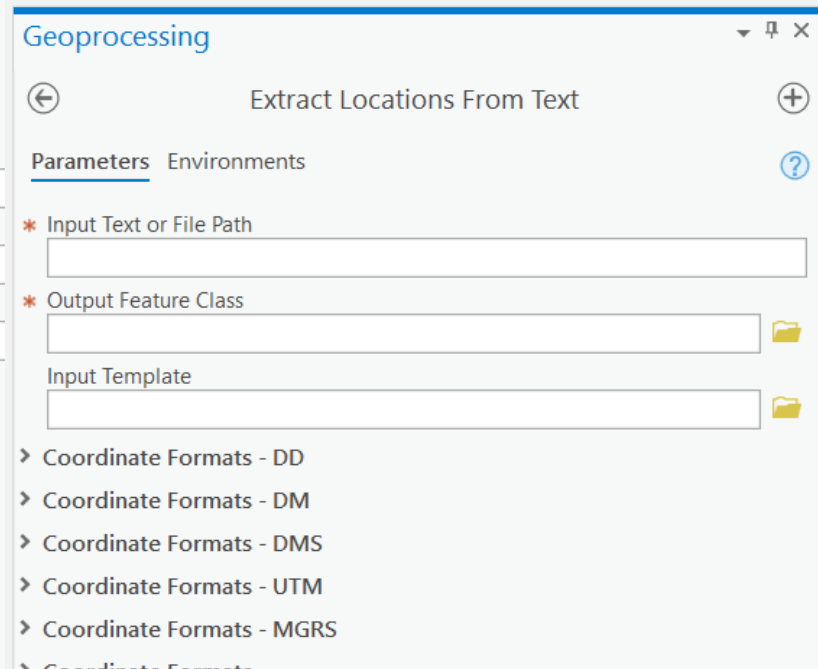
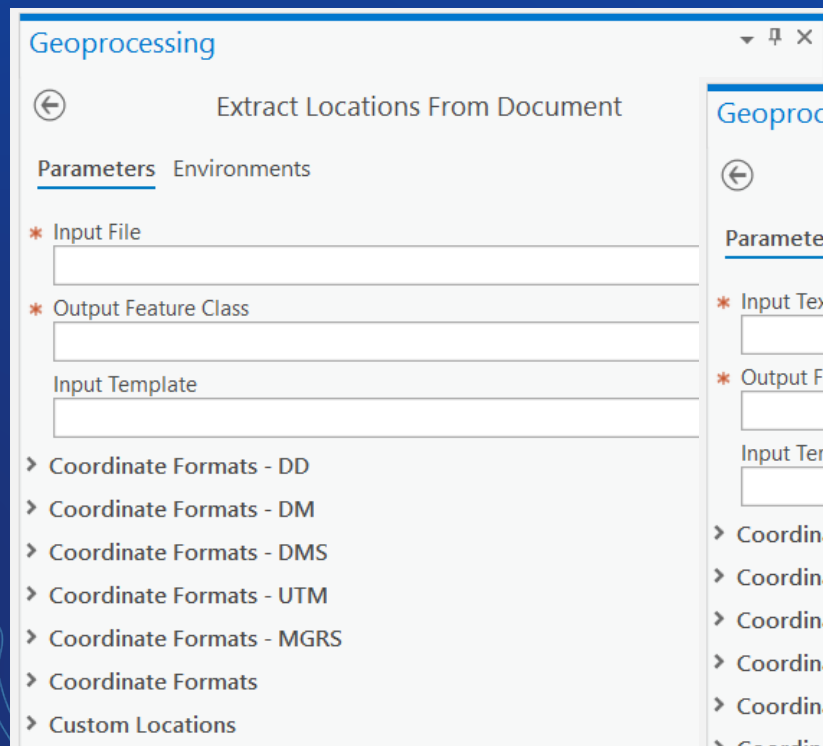
Extracting Locations in ArcGIS Pro

- 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)

Custom Attribute File

New File

Attributes

No attributes, add new, drop or import from another file.

Attribute Information

All active fields are required

Attribute Name	Field Name	Field Length
IED	IED	30

Search Options

Type: Near locations

Characters Before: 60

Characters After: 0

Matches: Keep all

Keywords

- *IED Capture only keyword
- *VBIED Capture only keyword
- *Explosive Device Capture only keyword

New keyword

Case Sensitive

Include in Capture

Capture Options

Capture Type: Capture only keyword

Number: 1

Case Sensitive

Include in Capture

Add Keyword Cancel

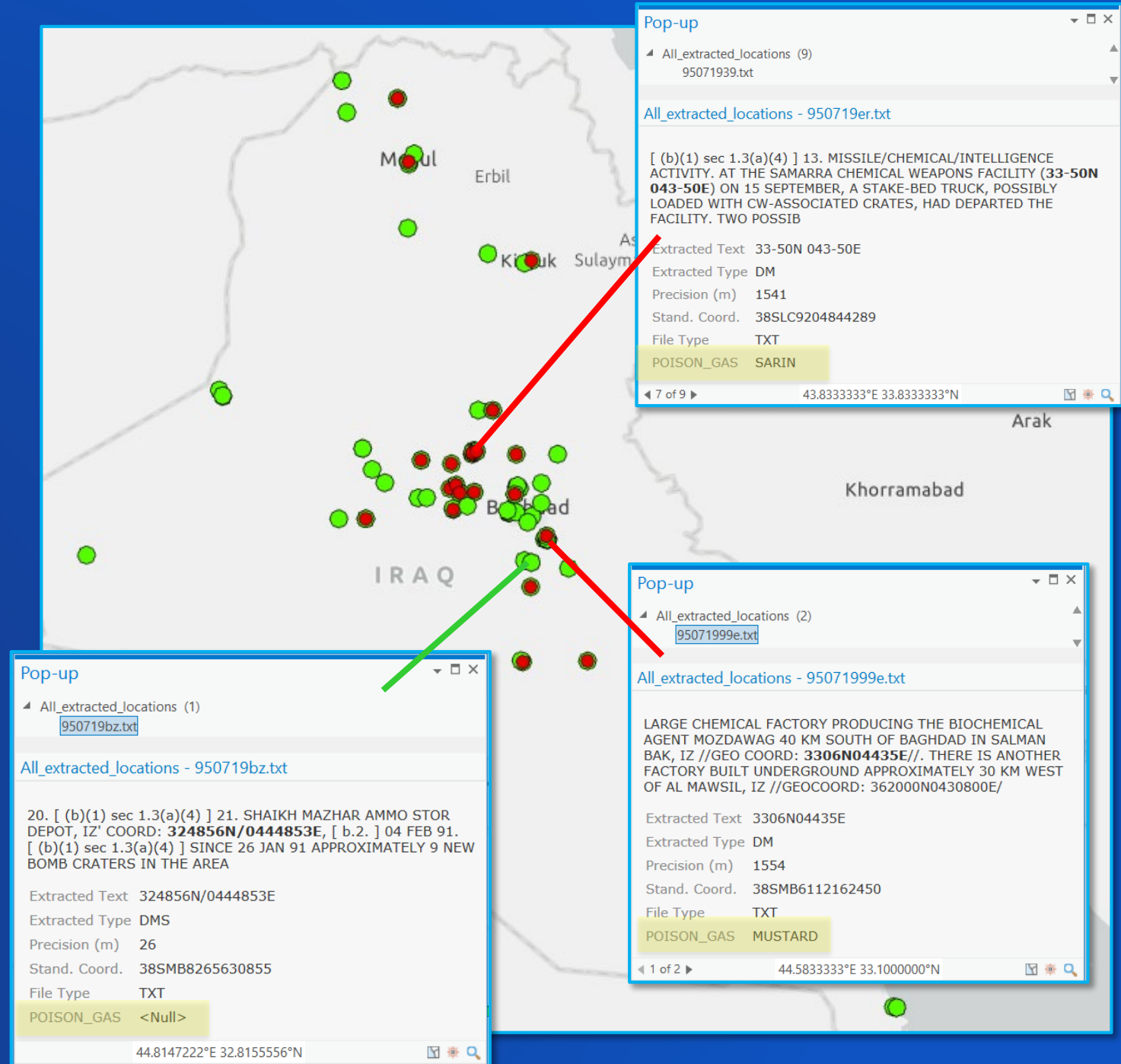
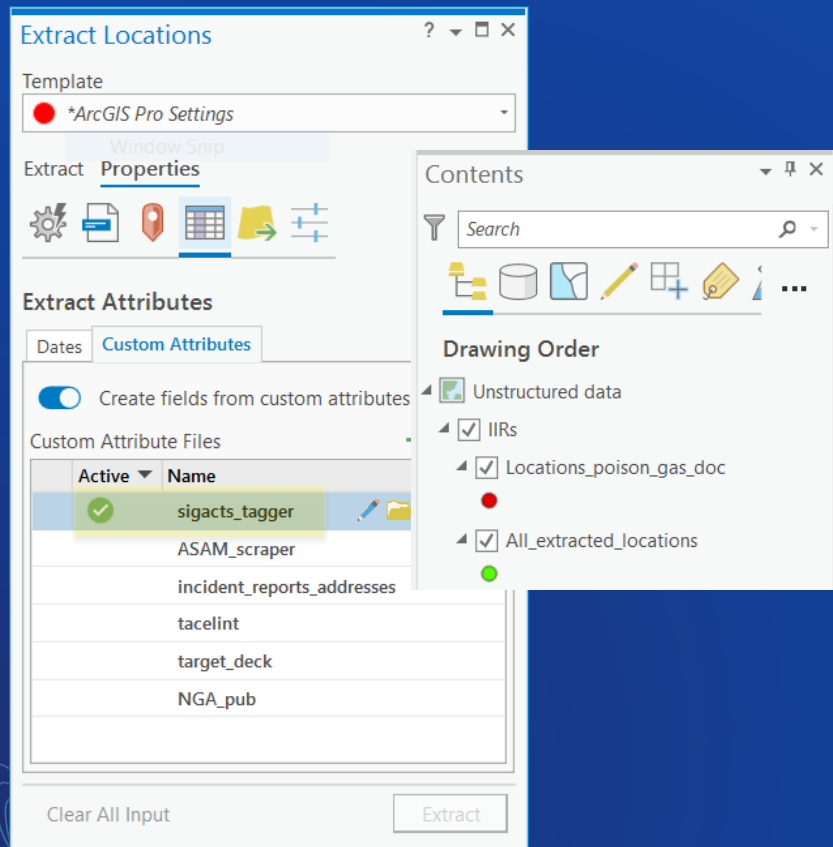
Add Attribute Cancel

Learn more about custom attributes

Save Close

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

Edit Custom Attribute

Name: Obstruction

Search: Near locations

GIS Name: Obstruct

GIS Length (chars): 254

Before (chars): 20

After (chars): 0

Matches: Keep only first

Keyword list:

Keyword	Description
buoy	Capture only keyword
light	Capture only keyword
seamount	Capture only keyword
rock	Capture only keyword

OK Cancel

Contents

Search

Drawing Order

- Unstructured data
 - Maritime
 - Possible_obstructions
 - All_extracted_locations

Pop-up

All_extracted_locations (8)

28°38'S, 16°27'E

All_extracted_locations - 28°38'S, 16°27'E.

Area No. 1, a coastal strip about 140 miles wide, extends S from latitude 26°S to the Orange River (28°38'S, 16°27'E). Landing or entry without permission is prohibited in this area. 5.18 Marine Mining Vessels (MMV),

Possible Obstruction: <Null>

Extracted Text 28°38'S, 16°27'E.

Extracted Type DM

Precision (m) 1628

Stand. Coord. 33JXJ4173431777

Obstruction <Null>

1 of 8 16.4500000°E 28.6333333°S

Pop-up

All_extracted_locations (2)

34°23'S, 18°30'E

All_extracted_locations - 34°23'S, 18°30'E.

extend about 1 mile SW from Cape Maclear and the sea generally breaks over them. 5.59 Bellows Rock (34°23'S, 18°30'E), which dries 1m, lies 2 miles SSW of Cape Point Light and the sea always breaks over it. The posit

Possible Obstruction: Rock

Extracted Text 34°23'S, 18°30'E.

Extracted Type DM

Precision (m) 1531

Stand. Coord. 34HBG7014692506

Obstruction Rock

1 of 2 18.5000000°E 34.3833333°S

Pop-up

All_extracted_locations (1)

31°41'S, 8°20'E

All_extracted_locations - 31°41'S, 8°20'E

h of the Orange River. A similar depth was reported (1985) to lie 5 miles further SE. Vema Seamount 31°41'S, 8°20'E. Reported depth of 7m Lies 465 miles WSW of the mouth of the Orange River and constitutes a danger t

Possible Obstruction: Seamount

Extracted Text 31°41'S, 8°20'E

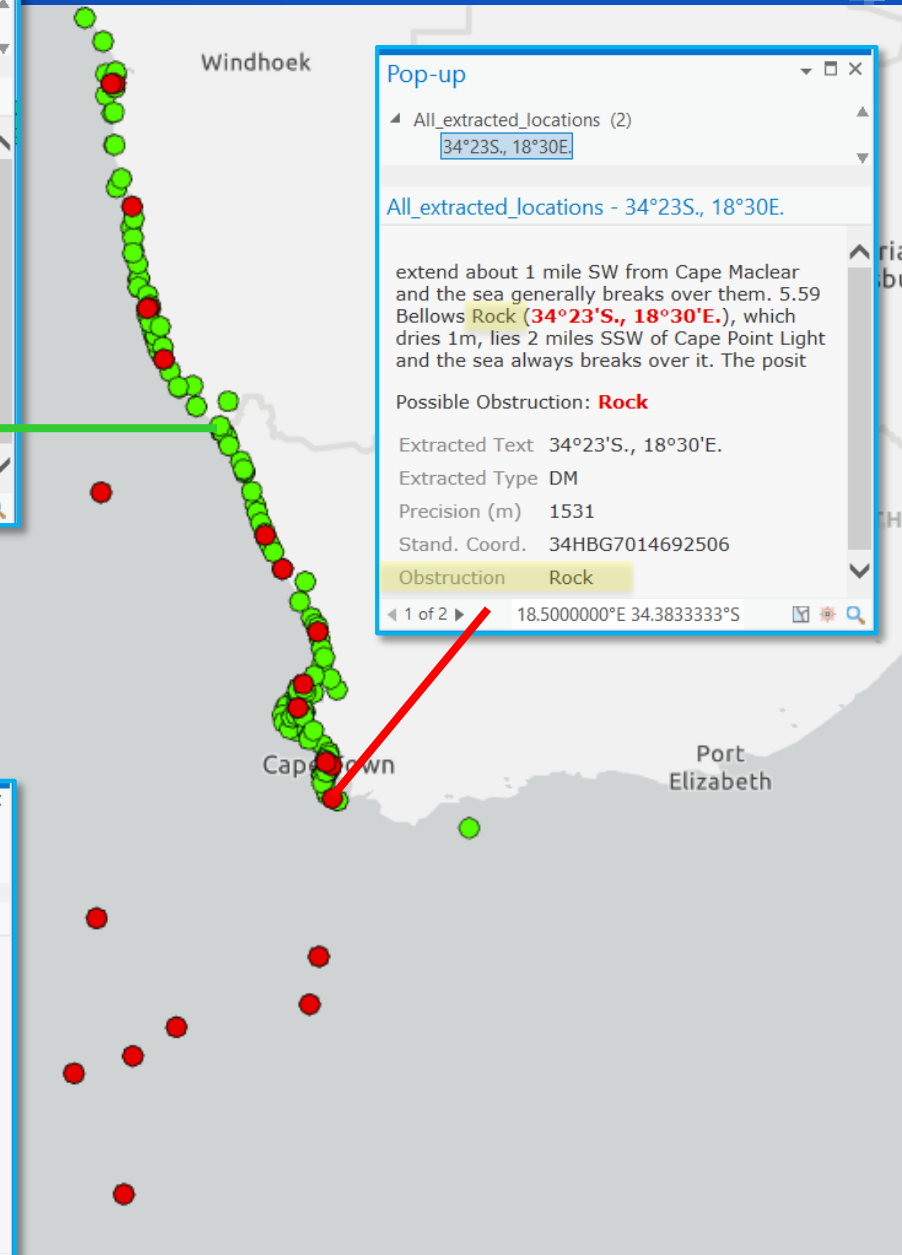
Extracted Type DM

Precision (m) 1579

Stand. Coord. 32JMK3681494470

Obstruction Seamount

8.3333333°E 31.6833333°S



Custom capture text based on keywords found in proximity to location

Edit Custom Attribute

Name

Aggressor

Search

Near locations

GIS Name

Aggressor

GIS Length (chars)

254

Before (chars)

0

After (chars)

50

Matches

Keep only first

Keyword list

Keyword

Aggressor:

Description

Capture until stop string

New

Edit

Remove

OK

Cancel

Edit Custom Attribute

Name

Description

Search

Near locations

GIS Name

Descript

GIS Length (chars)

254

Before (chars)

0

After (chars)

180

Matches

Keep only first

Keyword list

Keyword

Description:

Description

Capture until stop string

New

Edit

Remove

OK

Cancel

Edit Keyword

Keyword

Aggressor:

Capture control

Capture until stop string

Case sensitive

☐

Include in capture

☐

Stop string

Victim:

Case sensitive

☐

Include in capture

☐

OK

Cancel

- Coordinate Formats - DD
- Coordinate Formats - DM
 - Latitude And Longitude
 - X Y With Minutes Symbols
- Coordinate Formats - DMS
 - Latitude And Longitude
 - X Y With Seconds Symbols
 - X Y With Separators

Edit Keyword

Keyword

Description:

Capture control

Capture until stop string

Case sensitive

☐

Include in capture

☐

Stop string

Date of Occurrence:

Case sensitive

☐

Include in capture

☐

OK

Cancel

Maritime_incidents.txt - Notepad

File Edit Format View Help

Date of Occurrence: 10/29/2018

Reference Number: 2018-170

Geographical Subregion: 57

Geographical Location: 4° 46' 00" S

10° 08' 00" E

Aggressor: PIRATES

Victim: TANKER

Description: (U) REPUBLIC OF CONGO: On 29 October, pirates in a speed boat chased and fired upon LPG tanker BW FRIGG carrying out bunkering operations underway near position 04:46S - 010:08E, 101 nm, west of Pointe Noire. Emergency bunker stop and cast off initiated. Speed increased and the tanker made evasive maneuvers and escaped. All crew reported safe. (IMB; MDAT-GOG; www.pvilt.com; www.sguardian.com; www.fleetmon.com)

Date of Occurrence: 10/29/2018

Reference Number: 2018-169

Geographical Subregion: 57

Geographical Location: 4° 57' 00" S

10° 43' 00" E

Aggressor: ARMED PIRATES

Victim: SUPPLY VESSEL

Description: (U) REPUBLIC OF CONGO: On 29 October, offshore supply vessel ARK TZE was boarded and hijacked by armed pirates near position 04:57S - 010:43E, 68 nm west of Pointe Noire. All crew taken hostage and made to lie on deck while the pirates, ransacked and stole crew and ship's properties. The pirates kidnapped four crew and escaped. The remaining crew sailed the vessel to a safe port. One crewman reported injured. Bridge equipment damaged during the attack. (IMB; www.fleetmon.com)

Date of Occurrence: 10/26/2018

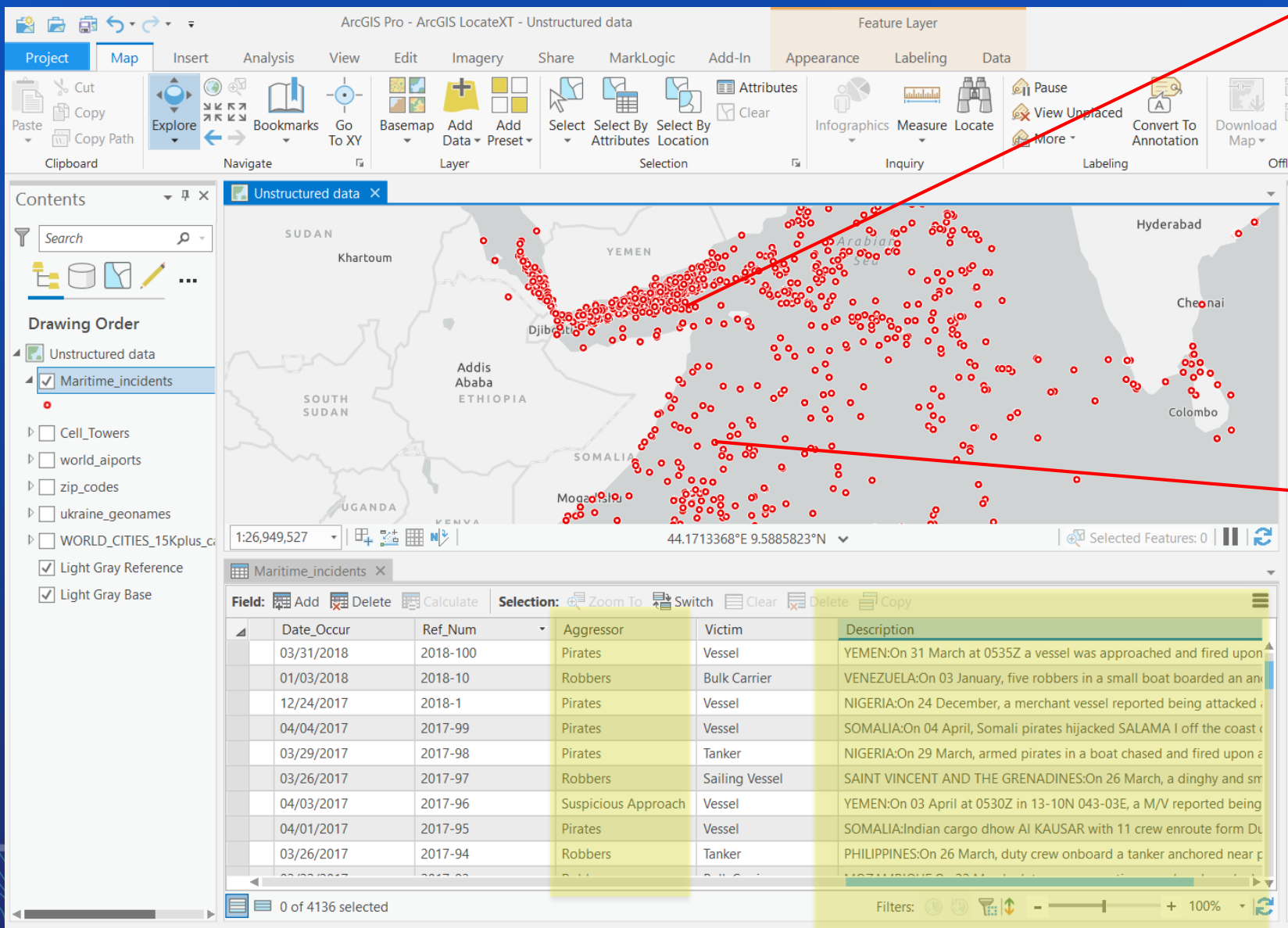
Reference Number: 2018-171

Geographical Subregion: 57

Geographical Location: 3° 40' 00" N

6° 40' 00" E

Custom capture text based on keywords found in proximity to location



Pop-up

Maritime_incidents (1)

Pirates

Maritime_incidents - Pirates

DESCRIPTION: INDIAN OCEAN:On 23 Feb, Singapore-flagged product tanker LEOPARD SUN was attacked by pirates, near 03-26N 050-27E, 159 nm off the central coast of Somalia. The tanker was approached by three skiffs and was fired upon by armed pirates. The armed security

Extracted Text 3° 26' 00" N 50° 27' 00" E

Extracted Type DMS

Precision (m) 31

Stand. Coord. 03.433333N 050.450000E

Date_Occur 02/23/2018

Ref_Num 2018-77

Aggressor Pirates

Victim Tanker

50.450000°E 3.433333°N

Pop-up

Maritime_incidents (1)

Suspicious Approach

Maritime_incidents - Suspicious Approach

DESCRIPTION: GULF OF ADEN:On 20 January, a small boat made a suspicious approach against a Panama-flagged merchant vessel under escort by a Chinese Navy frigate. The naval vessel fired warning flares and positioned itself for further action against the small boat, f

Extracted Text 11° 29' 00" N 44° 36' 00" E

Extracted Type DMS

Precision (m) 30

Stand. Coord. 11.483333N 044.600000E

Date_Occur 01/20/2018

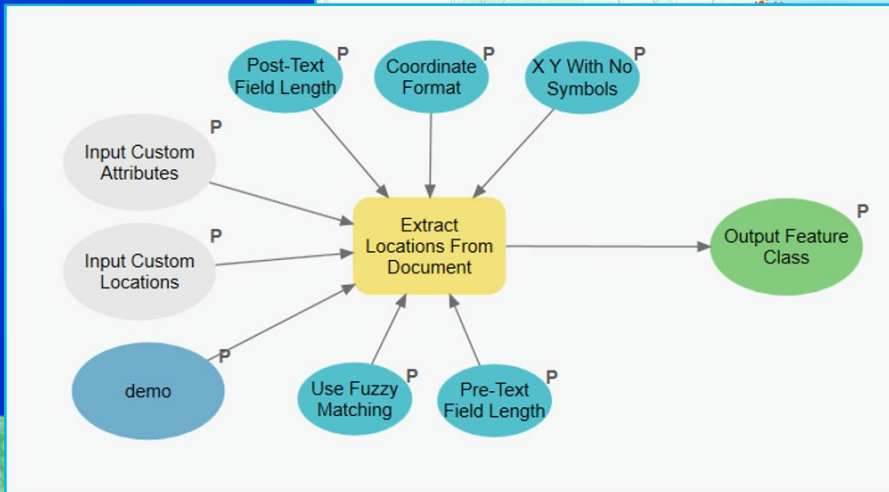
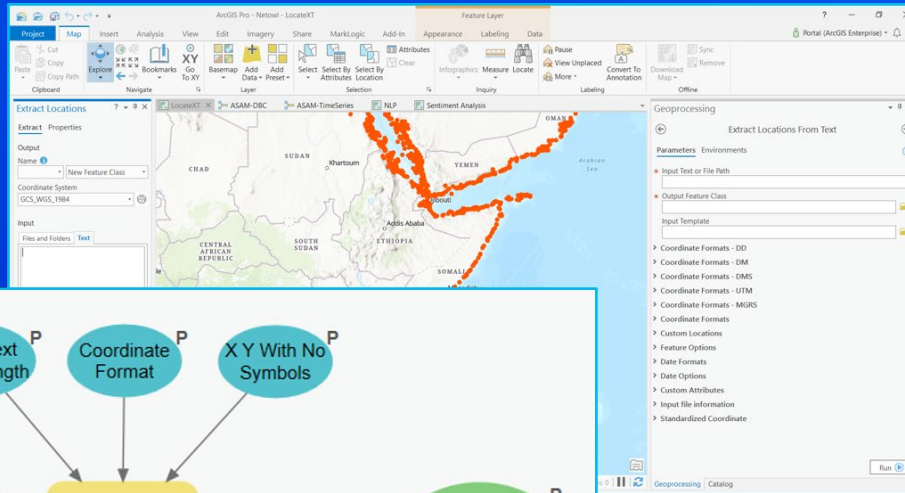
Ref_Num 2018-30

Aggressor Suspicious Approach

Victim Vessel

44.600000°E 11.483333°N

Explore Unstructured Data through LocateXT and Custom Attributes



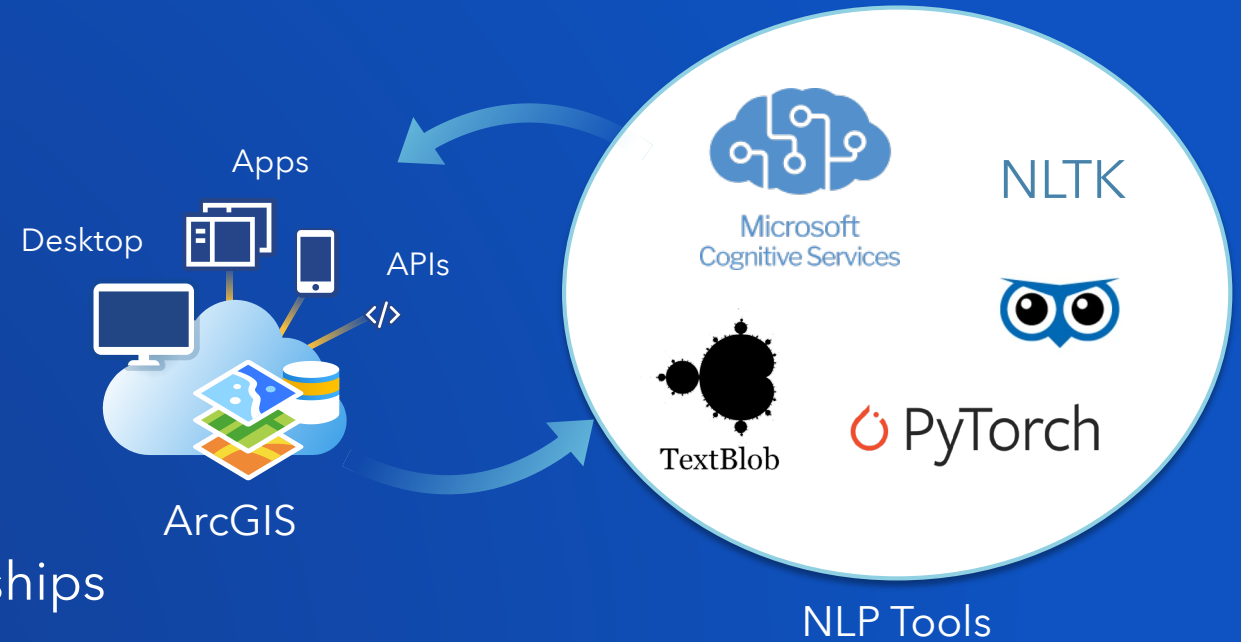
Integrating NLP

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

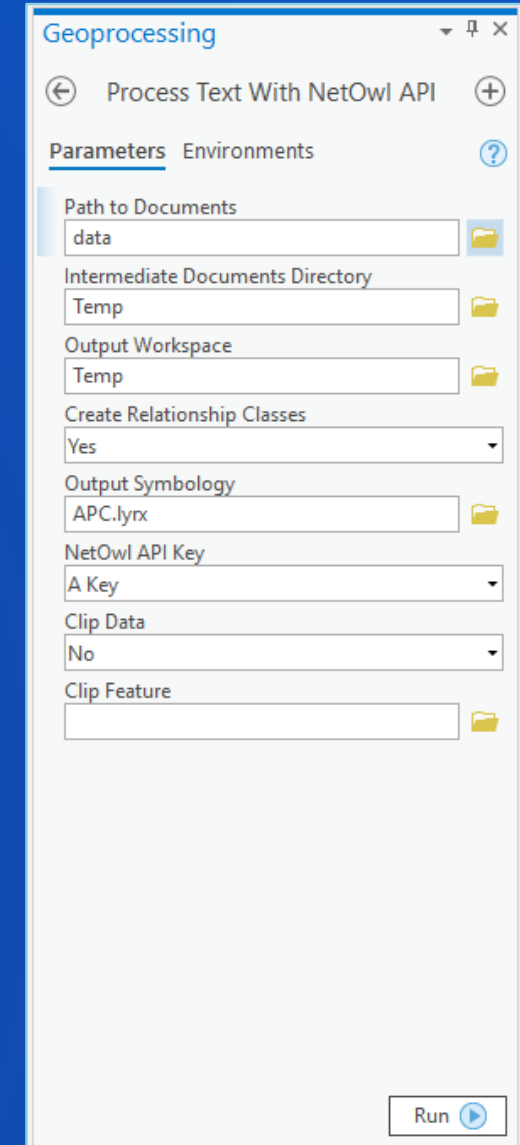
Possible Use Cases of Unstructured Data

- 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



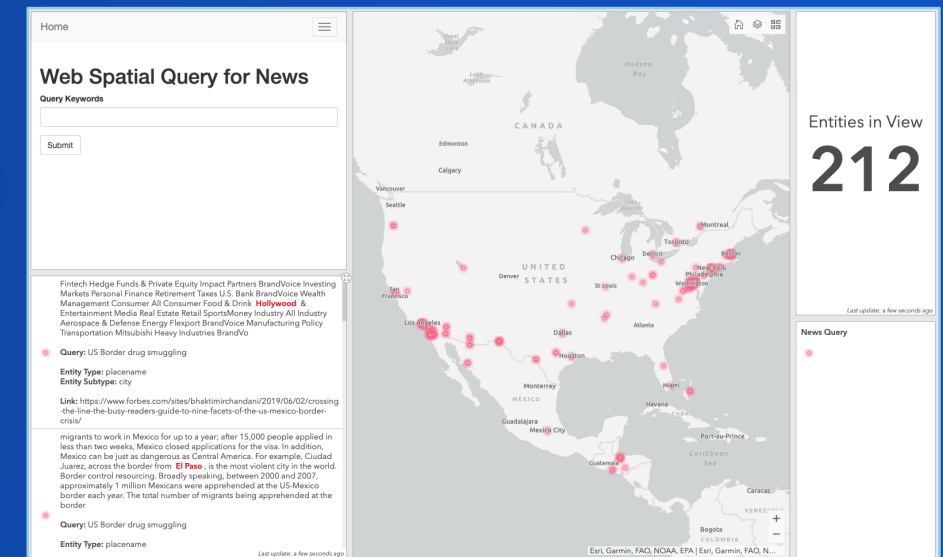
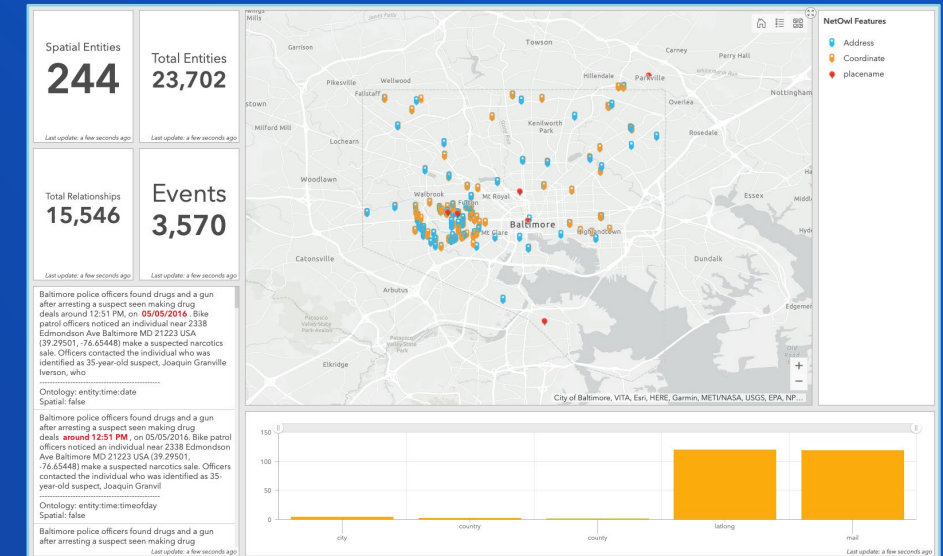
The screenshot shows the 'Geoprocessing' window in ArcGIS Pro. The title bar reads 'Geoprocessing'. Below the title bar, there are navigation icons (back, forward, search, etc.) and the tool name 'Process Text With NetOwl API'. The window is divided into two tabs: 'Parameters' (selected) and 'Environments'. The 'Parameters' tab contains the following fields:

- Path to Documents:** A text box containing 'data' with a folder icon to its right.
- Intermediate Documents Directory:** A text box containing 'Temp' with a folder icon to its right.
- Output Workspace:** A text box containing 'Temp' with a folder icon to its right.
- Create Relationship Classes:** A dropdown menu with 'Yes' selected.
- Output Symbology:** A text box containing 'APC.lyrx' with a folder icon to its right.
- NetOwl API Key:** A dropdown menu with 'A Key' selected.
- Clip Data:** A dropdown menu with 'No' selected.
- Clip Feature:** A text box with a folder icon to its right.

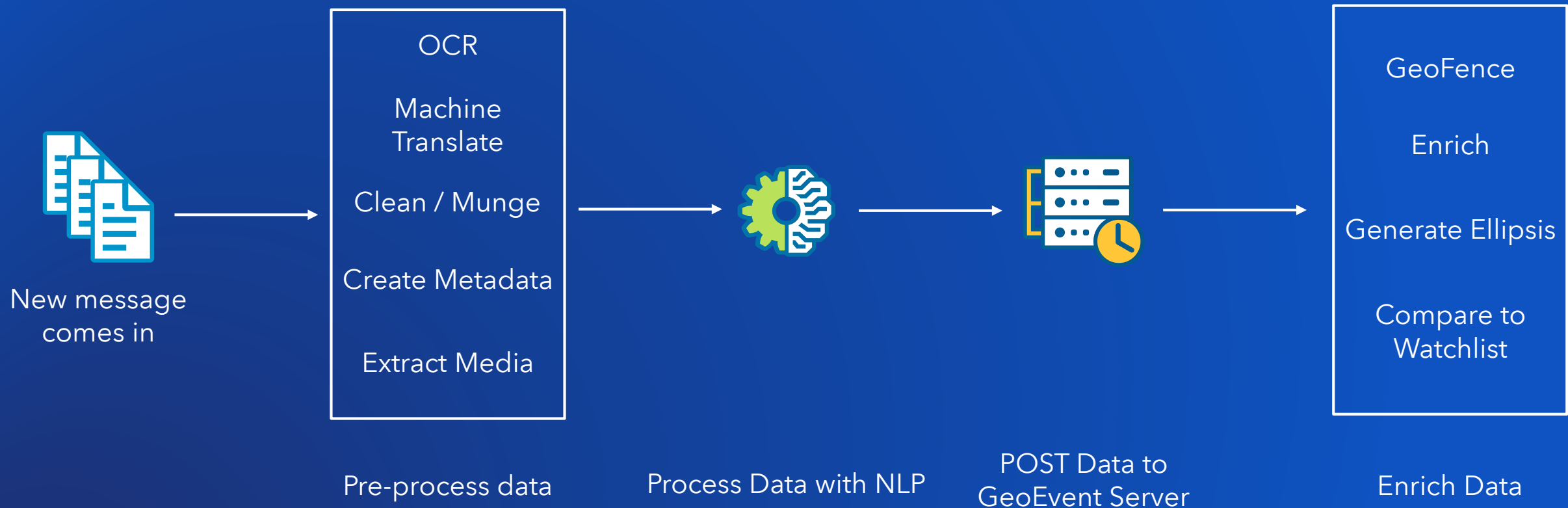
At the bottom right of the window is a 'Run' button with a play icon.

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.)

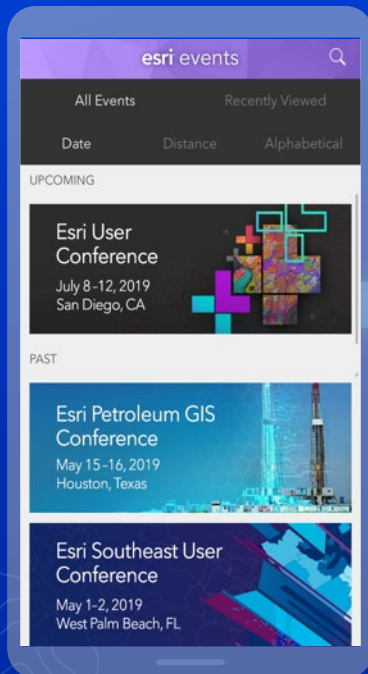


Building an Unstructured Data Pipeline

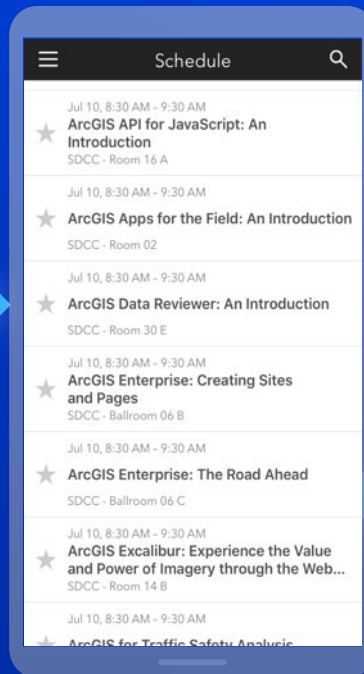


Please Share Your Feedback in the App

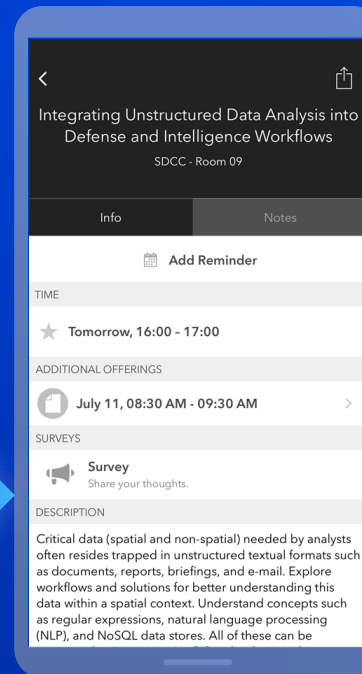
Download the Esri Events app and find your event



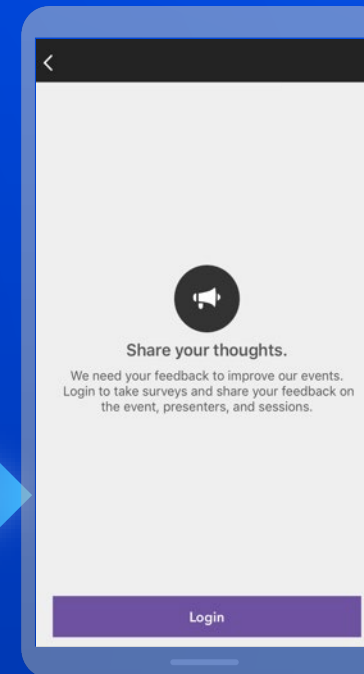
Select the session you attended



Scroll down to "Survey"



Log in to access the survey



Complete the survey and select "Submit"

