

5-2021

## SCOPE: Building and Testing an Integrated Manual-Automated Event Extraction Tool for Online Text-Based Media Sources

Matthew Crittenden

Follow this and additional works at: <https://scholarworks.wm.edu/honorsthesis>



Part of the [Databases and Information Systems Commons](#), and the [Data Science Commons](#)

---

### Recommended Citation

Crittenden, Matthew, "SCOPE: Building and Testing an Integrated Manual-Automated Event Extraction Tool for Online Text-Based Media Sources" (2021). *Undergraduate Honors Theses*. Paper 1651.  
<https://scholarworks.wm.edu/honorsthesis/1651>

This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact [scholarworks@wm.edu](mailto:scholarworks@wm.edu).

SCOPE: Building and Testing an Integrated Manual-Automated  
Event Extraction Tool for Online Text-Based Media Sources

A thesis submitted in partial fulfillment of the requirement  
for the degree of Bachelor of Arts in Data Science from  
William & Mary

by

Matthew Crittenden

Accepted for Honors



---

Dr. Daniel Runfola, Chair



---

Dr. Anthony Stefanidis



---

Dr. Maurits van der Veen

Williamsburg, VA  
May 12, 2021



WILLIAM & MARY  
CHARTERED 1693

# THE COLLEGE OF WILLIAM & MARY

## HONORS THESIS

---

**SCOPE: Building and Testing an Integrated Manual-Automated  
Event Extraction Tool for Online Text-Based Media Sources**

---

*Author:*

**Matthew CRITTENDEN**

*Advisor:*

**Dan RUNFOLA**

*A thesis submitted in fulfillment of the requirements for  
Interdisciplinary Honors in the degree of Bachelors of Arts in the*

**Data Science Program**

Accepted for Honors

---

Chair: Dr. Dan Runfola

---

Dr. Anthony Stefanidis

---

Dr. Maurits van der Veen

Williamsburg, Virginia

May 12, 2021

THE COLLEGE OF WILLIAM & MARY

# *Abstract*

Dr. Dan Runfola  
Data Science Program

Bachelors of Arts

## **SCOPE: Building and Testing an Integrated Manual-Automated Event Extraction Tool for Online Text-Based Media Sources**

by Matthew CRITTENDEN

Building on insights from two years of manually extracting events information from online news media, an interactive information extraction environment (IIEE) was developed. SCOPE, the Scientific Collection of Open-source Policy Evidence, is a Python Django-based tool divided across specialized modules for extracting structured events data from unstructured text. These modules are grouped into a flexible framework which enables the user to tailor the tool to meet their needs. Following principles of user-oriented learning for information extraction (IE), SCOPE offers an alternative approach to developing AI-assisted IE systems. In this piece, we detail the ongoing development of the SCOPE tool, present methods and results of tests of the efficacy of SCOPE relative to past methods, and provide a novel framework for future tests of AI-assisted IE tasks. Information gathered from a four-week period of use was analyzed to evaluate the initial utility of the tool and establish baseline accuracy metrics. Using the SCOPE tool, 15 users extracted 529 summaries and 362 structured events from 207 news articles achieving an accuracy of 31.8% holding time constant at 4 minutes per source. To demonstrate how fully or partially-automated AI processes can be integrated into SCOPE, a baseline AI was implemented and achieved 4.8% accuracy at 3.25 seconds per source. These results illustrate the ability of SCOPE to present the relative strengths and weaknesses of manual users and AI, as well as establish precedent and methods for integrating the two.

# Contents

<b>Abstract</b>	<b>i</b>
<b>1 Thesis</b>	<b>1</b>
1 Introduction	1
1.1 Literature Review Part 1: Existing IIEEs and Other IE Systems	3
1.1.1 Manual IE Systems	5
1.1.2 Automated IE Systems	7
1.1.3 Interactive IE Systems	8
1.1.4 User-oriented Learning IIEEs	9
1.2 Literature Review Part 2: Measuring Efficacy of IE Technologies	10
2 Methods	13
2.1 Introducing the SCOPE Method	14
2.2 Proceduralizing the Assessment of the Efficacy of IE Methods in SCOPE	17
2.3 Case Example: Developing a Comparative Efficacy Metric for Manual and Machine-Automated Event Extraction	20
2.3.1 Establishing a Baseline Using SCOPE for Manual Event Extracting and Parsing	20
2.3.2 Constructing Prototype Machine-Automated Functions	21
3 Results and Discussion	22
3.1 Conclusion	25
4 Acknowledgements	27
<b>A Main Appendix</b>	<b>29</b>

<b>B Designing the Scope Tool</b>	<b>37</b>
B.1 User-Oriented Learning-based System . . . . .	37
B.2 Modules, Classes, Suites, and Frameworks . . . . .	39
<b>C Graphics of the Scope Tool</b>	<b>47</b>
<b>References</b>	<b>61</b>

# List of Figures

1.1	Event Extraction from an Online News Article . . . . .	2
1.2	Images of geoParsing's Google Sheets IE Setup . . . . .	6
1.3	SCOPE's User-oriented Learning-based System . . . . .	14
1.4	Confusion Matrix for the Extracting Module . . . . .	18
1.5	Plot of F1 Score and Speed Attained by Each Method . . . . .	24
A.1	Source Worksheet of geoParsing's BRIGHT Google Sheet . . . . .	29
A.2	Activities Worksheet of geoParsing's BRIGHT Google Sheet . . . . .	30
A.3	Geocoding Worksheet of geoParsing's BRIGHT Google Sheet . . . . .	30
A.4	Source Worksheet of geoParsing's TRACAR Google Sheet . . . . .	31
A.5	Activities Worksheet of geoParsing's TRACAR Google Sheet . . . . .	31
A.6	Geocoding Worksheet of geoParsing's TRACAR Google Sheet . . . . .	32
A.7	Time Taken to Extract Information By Each Method . . . . .	32
A.8	Plot of Precision and Speed Attained by Each Method . . . . .	34
A.9	Plot of Precision and Speed Attained by Each Method (Stop Words) . . . . .	34
A.10	Plot of Recall and Speed Attained by Each Method . . . . .	35
A.11	Plot of Recall and Speed Attained by Each Method (Stop Words) . . . . .	35
A.12	Plot of F1-Score and Speed Attained by Each Method . . . . .	36
A.13	Plot of F1-Score and Speed Attained by Each Method (Stop Words) . . . . .	36
B.1	SCOPE's User-oriented Learning-based System (repeat) . . . . .	37
B.2	Diagram of SCOPE's Modules in a Comprehensive Framework . . . . .	40

# List of Tables

1.1	Different IE Systems . . . . .	5
1.2	Accuracy Statistics for Each Method . . . . .	22
A.1	Extract Counts by Source . . . . .	33
A.2	Accuracy Statistics in Terms of False Positives, False Negatives, True Positives, and True Negatives . . . . .	33
A.3	Accuracy Statistics for Each Method (repeat) . . . . .	34



# List of Abbreviations

AA	Auto-Assisted
ACLED	Armed Conflict Location & Events Data
AI	Artificial Intelligence
CSV	Comma Separated Values
EE	Event Extraction
GDELT	Global Dataset of Events Locations and Tone
geoLab	Geospatial Evaluation & Observation Lab
IE	Information Extraction
IIE	Interactive Information Extraction
IIEE	Interactive Information Extraction Environment
ML	Machine Learning
NLP	Natural Language Processing
QA	Quality Assurance
SCOPE	Scientific Collection of Open-source Policy Evidence
SW	Stop Words
TW	Trigger Words

# Chapter 1

## Thesis

### 1 Introduction

The amount of online text-based data continues to increase at a rate which exceeds our processing capabilities - both human and, today, computational (Wang, 2017). The creation of global events datasets such as GDELT and ACLED showcases the value and utility of artificial intelligence (AI) for developing efficient methods of extracting information from these massive online sources.<sup>1</sup> Automated extraction of specific information from open-source texts has become a popular area of research (Wang, 2017; Naughton, Kushmerick, and Carthy, 2006; Tong et al., 2020; Goswami and Kumar, 2016; Ritter, Etzioni, and Clark, 2012; Salam et al., 2018). Event extraction (EE), a field of information extraction (IE), seeks to “detect event instance(s) in texts, and if existing, identify the event type as well as all of its participants and attributes” (Xiang and Wang, 2019, 2). Through event extraction, we are able to trim down pages and paragraphs of details to the key 5W1H (who, what, when, where, why, and how) facts of an event (Xiang and Wang, 2019; Chan et al., 2019). For the purposes of this thesis, we define an event as a “specific occurrence of something that happens in a certain time and a certain place involving one or more participants, which can frequently be described as a change of state” (Xiang and Wang, 2019, 2).

While existing global events datasets are valuable for identifying trends at aggregate, they are often either domain-specific (e.g., ACLED) or too broad (e.g.,

---

<sup>1</sup>GDELT is a realtime, open-source network and database of global human society which monitors the world’s print, broadcast, and web news in over 100 languages. <https://www.gdeltproject.org/>; ACLED collects real-time data on the locations, dates, actors, fatalities, and types of all reported political violence and protest events in most regions of the world. <https://acleddata.com/>.

GDELT) to be of use in smaller-scope projects. Furthermore, the construction of an events dataset using AI traditionally requires expertise in natural language processing (NLP) methods and domain-specific linguistic patterns (Cardie and Pierce, 1998). As such, existing events datasets do not always meet the needs of researchers and the barriers to entry for creating novel events datasets are high.

On the other hand, constructing novel events datasets by hand is also a tedious and time-intensive process (Chen et al., 2017; Halterman et al., 2017; Duan, He, and Zhao, 2017). This thesis was conducted in the context of constructing events datasets on Chinese Belt and Road Initiative development projects in Latin America and the Caribbean and Russian foreign relations in the Central African region. The team of students collecting this data - geoParsing - follows a three-step process of identifying sources of relevant information, parsing that information into spreadsheets, and quality assuring both stages. Figure 1.1 provides an example of event extraction from an online news article, in which text describing an event - the building of a new hospital block - is used to inform the coding of fields for the structured data (i.e., what took place, who was involved, where and when did the event happen, and so on).<sup>2</sup>

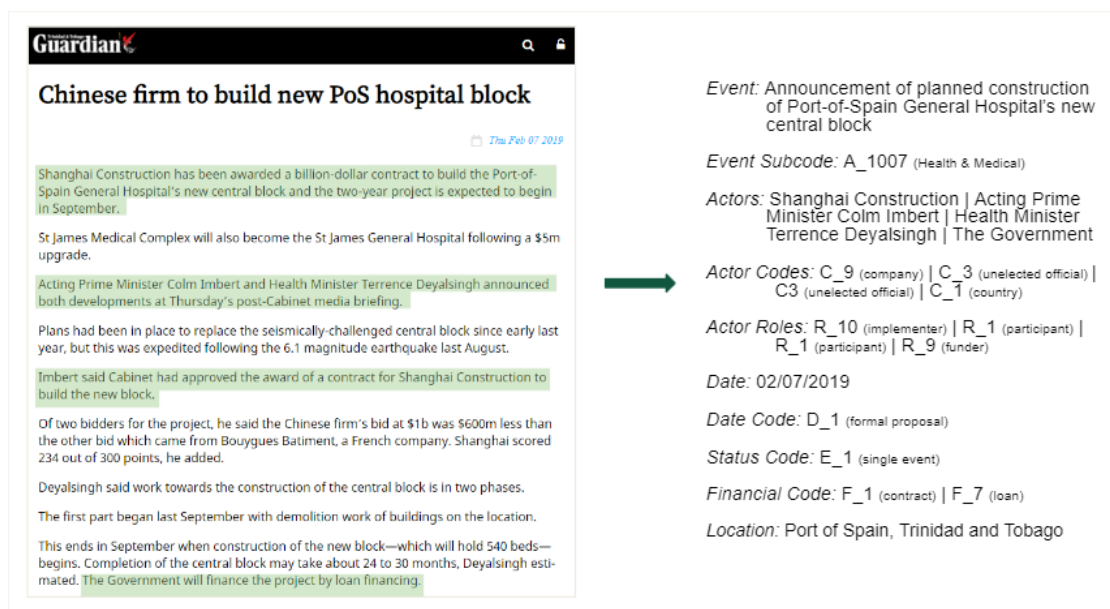


FIGURE 1.1: Event Extraction from an Online News Article

<sup>2</sup><https://www.guardian.co.tt/news/chinese-firm-to-build-new-pos-hospital-block-6.2.775571.0e767b4a5f>

Over a two-year period, geoParsing has amassed about 400 observations parsed from over 1,200 online sources (primarily news articles).<sup>3</sup> These observations possess meticulous attention to detail and have been subjected to several rounds of quality assurance, but their quantity is nonetheless alarmingly low, and reflective of the unsustainable time costs of purely manual approaches to event extraction.

Recognizing these challenges, this paper presents the Scientific Collection of Open-source Policy Evidence (SCOPE), an interactive information extraction environment (IIEE) which enables researchers to develop their own structured events datasets from unstructured text data. Section 1.1 provides an overview of existing IEs and their respective strengths and weaknesses. Section 1.2 discusses how existing information extraction technologies measure the efficacy of their methods. Section 2 details the design and implementation of the SCOPE tool, methods for comparing the efficacy between different information extraction methods within SCOPE, and a case example establishing a baseline for comparisons between manual and preliminary automated methods of event extraction. Section 3 provides a discussion and conclusion, including the results of the case example and plans for future improvements.

## 1.1 Literature Review Part 1: Existing IIEEs and Other IE Systems

Cardie and Pierce first proposed the development of an interactive information extraction environment (IIEE) in 1998 to allow end-users the ability to create and use information extraction methods without the need for specialized knowledge in NLP and computational linguistics (Cardie and Pierce, 1998). In their words, an IE system is more broadly “a natural language processing system that produces structured data summaries of input text” (Cardie and Pierce, 1998, 1). IE systems are used in many domains – from counter-terrorism to medicine –

---

<sup>3</sup>The author founded the geoParsing Team in October 2018 and assembled a group of 4 other undergraduate students to begin working in January 2019. Since then, the team has trained about 30 researchers, with between 10-15 researchers working each semester on data collection in addition to other tasks, such as analyzing the data and writing reports. The team’s work is hosted on <http://www.scopedata.org/> for free, public, open-source use. This data has gone on to be used in multiple publications on our website, <http://www.scopedata.org/Reports.php>, and with partner organizations in the U.S. Government at <https://www.tearline.mil/>.

to perform tasks including identifying source documents, presenting their information in easily digestible ways, and automatically filtering and structuring information (ACLED, 2019; GDELT, 2015; Pafilis et al., 2016; Cejuela et al., 2014). The traditional methods for developing an IE involve the annotation of hundreds of documents in a large training corpus, domain-specific knowledge engineering, and expertise in NLP system design (Cardie and Pierce, 1998; Halterman et al., 2017; Sarawagi, 2007). An IIEE is proposed to mitigate these costly and highly technical barriers.

In an IIEE, a user would interact directly with the IE system to extract whatever information they want, code it to a schema of their own design, and train a machine over time to replicate this process (Pierce and Cardie, 2001; Culotta et al., 2006). This process is described as user-oriented learning, a “method of developing IE systems which recognizes the complementary strengths of the human user and the IE system” (Pierce and Cardie, 2001, 1-2). This process is also described as corrective feedback and persistent learning, where a user makes corrections to an automated machine-learned output and the model continually learns from these corrections (Culotta et al., 2006). The more efficient the corrective feedback mechanisms are, the more effective persistent learning will be. This method of user-oriented learning could be even more efficacious for IE today given the invention of deep learning techniques which could grow alongside users’ interactions with the extraction environment (Chan et al., 2019). In the following sections, I will introduce how IEs – interactive and not – have been utilized since Cardie and Pierce’s proposal more than two decades ago. This discussion will enable us to situate the development and use of the SCOPE tool within the context of other existing IE technologies, many of which are found in the social sciences and biomedical field. Table 1.1 provides examples of different types of IE systems.

Type of IE System	Example
Manual IE	geoParsing, ACLED (ACLED, 2019; ACLED, 2020)
Automated IE	GDELT (GDELT, 2015; National Statistics, 2020; Wang, 2017), Biryani (Halterman et al., 2017), Proteus BIO (Grishman, Huttunen, and Yangarber, 2002)
Interactive IE	Egos (Campos et al., 2014), EXTRACT (Pafilis et al., 2016)
User-oriented Learning IIEE	tagtog (Cejuela et al., 2014), SCOPE

TABLE 1.1: Different IE Systems

### 1.1.1 Manual IE Systems

Today, many manual IE workflows are contained with spreadsheets functioning as manual event extraction environments. In a spreadsheet, the 5W1H information of an event makes up the columns and each event observation is a row. As an example from our own substantive work, this information would relate to the domains of Chinese and Russian foreign policy events. Fields include source information, the names and types of actors, actions, dates, geocoded locations, and other details describing the event, which we refer to synonymously as the activity (see Figure 1.2)). The information collected is consistent with that of other political events datasets (Halterman et al., 2017; Salam et al., 2018; ACLED, 2019). See Appendix A for larger images of the geoParsing Google Sheets.

A spreadsheet-based approach to data entry is widely used and can be adapted to event extraction (Broman and Woo, 2018; Taylor et al., 2020). Most spreadsheet programs, such as Google Sheets or Excel, allow users to create separate worksheets for each step in the event extraction workflow (e.g., finding sources, parsing them for events, geocoding the events) and perform the basic tasks for data entry, storage, analysis, and visualization (Broman and Woo, 2018; Taylor et al., 2020). However, current spreadsheet programs are error prone and do not offer a robustness to user errors (e.g., typos, working simultaneously on the same row, accidentally deleting cell contents) which is vital to user-oriented

The figure displays six screenshots of Google Sheets spreadsheets, arranged in a 3x2 grid. Each screenshot shows a different sheet within a spreadsheet application, likely Google Sheets. The sheets are labeled as follows:

- BRIGHT Sources:** Shows a table with columns for Source ID, Source Name, and other metadata. The data rows are highlighted in green.
- TRACAR Sources:** Shows a similar table to BRIGHT Sources, but with a prominent pink vertical highlight on one of the columns.
- BRIGHT Activities:** Shows a table with columns for Activity ID, Activity Name, and other details. The data rows are highlighted in green.
- TRACAR Activities:** Shows a table with columns for Activity ID, Activity Name, and other details. The data rows are highlighted in green.
- BRIGHT Geocoding:** Shows a table with columns for Activity ID, Activity Name, and other details. The data rows are highlighted in green.
- TRACAR Geocoding:** Shows a table with columns for Activity ID, Activity Name, and other details. The data rows are highlighted in green.

FIGURE 1.2: Images of geoParsing’s Google Sheets IE Setup

learning IEs, given the value of manually created training data to develop automated components (Broman and Woo, 2018; Taylor et al., 2020). Spreadsheet programs also do not offer the necessary infrastructure to develop and implement machine learning models to integrate into the event extraction process – experts recommend that spreadsheets are best suited for data entry and storage, and analysis should be conducted separately (Broman and Woo, 2018). As such, the methodological improvements attainable in a spreadsheet-based IE system are limited.

Another known manual IE system is the methodology which underlies ACLED, the Armed Conflict Location & Event Data Project (ACLED, 2019). ACLED is a “disaggregated data collection, analysis, and crisis mapping project” which collects “real-time data on the locations, dates, actors, fatalities, and types of all reported political violence and protest events” in several regions of the world

(ACLED, 2021). ACLED describes its methodology as one in which every week researchers manually assess thousands of sources to extract relevant information following rules on who, what, when, where, and when (ACLED, 2019; ACLED, 2020). Extracted information undergoes at least three rounds of review to ensure validity through intra- and inter-coder checks as well as correcting other user errors in coding (ACLED, 2019).

### 1.1.2 Automated IE Systems

GDELT, the Global Dataset of Events Location and Tone Project, publishes events data automatically extracted from online news media around the world (GDELT, 2015). GDELT is supported by Google Jigsaw and Google BigQuery to use machine learning to extract information from sources every 15 minutes, resulting in hundreds of thousands of rows per day (GDELT, 2015). Overall, there is limited public documentation of the steps in the automated process GDELT uses to extract and aggregate information from individual news articles into events, making it difficult to reconstruct here (National Statistics, 2020; Wang, 2017). It is also unclear whether specific quality assurance mechanisms are built into GDELT, as they are in the much smaller-scope projects such as ACLED (ACLED, 2019). Wang identifies two weaknesses of GDELT in the abundance of duplicate events and overrepresentation of a few domains across the majority of events (Wang, 2017). Further, in an assessment of a random sample of 3,000 articles from GDELT between March and May 2014, Wang found that GDELT achieved an average event coding accuracy of 16.2% when compared with at least one human coder (Wang, 2017). As such, given the scale and speed of GDELT, it seems that it follows a completely automated process which does not incorporate manual user elements and should still be regarded as experimental given significant concerns about the data quality (National Statistics, 2020; Wang, 2017). Wang suggests that introducing some human supervision could improve accuracy, albeit at a slight loss to the speed of a fully automated IE system (Wang, 2017).

Other automated IE systems include Biryani, a scalable system for extracting political events data, and Proteus BIO, which was designed to maintain a real-time database of infectious disease outbreaks (Halterman et al., 2017; Grishman, Huttunen, and Yangarber, 2002). Biryani is a modular, containerized system



using a Spark architecture for distributed CoreNLP processing and event extraction which allows social scientists to rapidly process sources across one or more machines (Halterman et al., 2017). Though Biryani is not interactive, it was still designed to be easily usable by social scientists, with its architecture built on Docker containers allowing dependencies and software to be self-contained (Halterman et al., 2017). Proteus BIO is also a modular system, but the IE tasks - not the processing capability - are split across modules (Grishman, Huttunen, and Yangarber, 2002). The first module is a crawler which identifies sources of information from the web (Grishman, Huttunen, and Yangarber, 2002). The second is an automated extraction engine which begins with tokenization and lexical look-up, then pattern matching to recognize actor and location names before adding observations to a database (Grishman, Huttunen, and Yangarber, 2002). The database is viewable through the third module, a web-based interface which displays the data as a spreadsheet, but is not editable given the system's automated design (Grishman, Huttunen, and Yangarber, 2002).

### 1.1.3 Interactive IE Systems

Interactive IE (IIE) systems are common in the biomedical field since curating some genomic resources could take decades and automated processes are limited by the complexity of the field (Campos et al., 2014). IIE systems offer a compromise between speed and accuracy. One IIE system which stands out is Egos, a web-based platform for text mining and curation which allows users to manually and automatically annotate documents (Campos et al., 2014). Although it does not integrate user-oriented learning elements, Egos was developed with focus on usability and simplicity, with features to establish annotation guidelines and user accesses, import documents, annotate interactively, and export documents. (Campos et al., 2014). The automated annotating function utilizes a REST API from the existing literature, the BeCASE REST API, to identify genes, proteins, species, chemicals, and other features of interest (Campos et al., 2014).

Another IIE system in the biomedical fields is EXTRACT, an interactive annotation tool which helps researchers identify scientific terms for annotation (Pafilis et al., 2016). EXTRACT works primarily as a browser bookmark which users click to classify text into scientific terms using named entity recognition (NER). Four previously published NER systems are integrated on the server's

backend to do flexible matching of a dictionary of millions of names against thousands of abstracts per second (Pafilis et al., 2016). A supervised approach has the user select a section or all of the text in a web page or document to be classified and annotated (Pafilis et al., 2016). All of the annotations are collected in tabular form with references to the location of the original text in the source document (Pafilis et al., 2016). There does not appear to be a user-oriented learning component to EXTRACT; rather, it is a tool which is meant to aid researchers and any improvements to its annotating ability would need to be made either to its user interface or to the NER systems which it references.

#### 1.1.4 User-oriented Learning IIEEs

Similar to the Egos and EXTRACT systems described above, the “tagtog” system is a web-based annotation framework used to mark up entities and concepts in full-text articles (Cejuela et al., 2014). The main difference is that tagtog integrated user-oriented learning to improve its results over time (Cejuela et al., 2014). tagtog leverages manual user annotation and machine-learned annotation to identify and extract gene symbols and names. The system uses a general-purpose named entity recognizer implemented with conditional random fields which result in a slightly lower performance than similar methods but have the benefit of increased speed, which is vital to a user-interactive application (Cejuela et al., 2014). Initially, the tool is trained with a small set of manually annotated documents. Then, as researchers work in the tagtog environment, sets of documents are automatically annotated for the researchers to review and validate. The interaction between the automated machine learning system and user feedback within tagtog allows for continuous and iterative retraining of the machine learning methods which can lead to an ever-improving performance in automatic prediction (Culotta et al., 2006).

Chan et al present a method for improving event detection by improving triggers and event types through user-oriented learning (Chan et al., 2019). They demonstrate that with less than 10 minutes of human effort per event type, their system achieved better performance for 67 novel event types when building from the basic ACE annotation dataset (Chan et al., 2019). Since machine learning algorithms are rarely perfect, they must be compensated for by interacting effectively with the user and the environment (Culotta et al., 2006).

## 1.2 Literature Review Part 2: Measuring Efficacy of IE Technologies

Measuring the efficacy of an IE system differs based on the design and methods of the system. For example, a user-oriented learning IIEE will consider metrics which differ, even if slightly, from those of a traditional IE system built using a corpus of annotated documents. In this section, a wide range of metrics are presented to later be implemented in the SCOPE tool.

There are three measurements of performance considered in the majority of event extraction literature: precision, recall, and F1 Score (Ahn, 2006; Campos et al., 2014; Culotta et al., 2006; Hogenboom et al., 2016; Pafilis et al., 2016). These metrics vary based on the task they are describing, which is in turn dependent on the IE system design. Most systems follow either pipeline classification or joint classification. Pipeline classification describes a process in which a set of independent classifiers are “each trained to complete one subtask and the output of one classifier can also serve as a part of the input to its successive classifier” (Xiang and Wang, 2019, 9). Joint classification describes a process in which the acts of identifying and parsing events are handled simultaneously (Xiang and Wang, 2019). As such, in joint classification, precision refers to the number of events which were properly extracted divided by the sum of the events which were properly extracted and the events which were extracted but should not have been (See Equation 1.1). The latter may not even constitute events. I.e., precision is the fraction of properly retrieved events which are relevant (Hogenboom et al., 2016). Recall is the number of events which were properly extracted divided by the sum of the events which were properly extracted and the events which were not extracted but should have been (See Equation 1.2). I.e., the fraction of relevant events that are properly retrieved (Hogenboom et al., 2016). Conversely, in pipeline classification, each of these measures refers to one step in the event extraction process, such as identifying event triggers (See Equations 1.3 and 1.4). F1 Score, the harmonic mean, is a balance of precision and recall (See Equation 1.5). It is more difficult to achieve a high recall, which requires knowledge of all the missed events, than a higher precision, which only requires knowledge of which extracted events should not be included (Sarawagi, 2007).

Joint Classification:

$$Precision = \frac{Events\ properly\ extracted}{Events\ properly\ extracted + Events\ extracted\ which\ should\ not\ have\ been} \quad (1.1)$$

$$Recall = \frac{Events\ properly\ extracted}{Events\ properly\ extracted + Events\ not\ extracted\ which\ should\ have\ been} \quad (1.2)$$

Pipeline Classification:

$$Precision = \frac{Events\ triggers\ properly\ identified}{Event\ triggers\ properly\ identified + Event\ triggers\ identified\ which\ should\ not\ have\ been} \quad (1.3)$$

$$Recall = \frac{Events\ triggers\ properly\ identified}{Event\ triggers\ properly\ identified + Event\ triggers\ not\ identified\ which\ should\ have\ been} \quad (1.4)$$

F1 Score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1.5)$$

Yet, these metrics only consider accuracy – not the other dimensions of performance in an IE system. Most event extraction literature – user-oriented learning-based and not – also tracks the time required to execute a task, either in part (pipeline classification) or fully (joint classification) (Campos et al., 2014; Culotta et al., 2006; Hogenboom et al., 2016; Pafilis et al., 2016). For example, Pafilis et al. find that their EXTRACT tool sped up the annotation process by some 15-25% (Pafilis et al., 2016). Campos et al. similarly find that their Egos tool reduced curation times by some 1.5 to 4 times (Campos et al., 2014). Culotta et al. also consider measures of confidence in event extraction using a variant of the sum-product, or forward-backward, algorithm (Culotta et al., 2006). Among metrics, Sarawagi identifies confidence as especially difficult to estimate from typical extraction models (Sarawagi, 2007). In any IE system, improvements to the system are understood as increases in precision, recall, F1 Score, and confidence or decreases in performance time (Ahn, 2006; Campos et al., 2014; Culotta et al., 2006; Hogenboom et al., 2016; Pafilis et al., 2016). This is not any different

from comparing the performance of different machine learning algorithms.

There are additional metrics, though, for interactive IE systems. In their discussion of EXTRACT, Pafilis et al. go on to evaluate their system by usability (Pafilis et al., 2016). Usability includes considerations of how easily navigable the tool or environment is and whether the user can accomplish the task (Pafilis et al., 2016). Campos et al. define usability for a system as having “easy-to-understand interfaces, and simple installation and configuration steps” (Campos et al., 2014, 3). Usability is usually a qualitative metric, and is transformed into a quantitative metric via a binary (e.g., “Yes, the system was helpful and easy to navigate” = 1 and “No, the system was not helpful” = 0) or a scale (e.g., “rate the system’s helpfulness from 1 to 5, with 5 being the most helpful”) (Pafilis et al., 2016; Campos et al., 2014). Huttunen et al. propose to measure the relevance of an extracted event to the user regardless of whether the event was properly extracted or not (Huttunen et al., 2013). In other words, if the user is constructing a dataset on Chinese Belt and Road Initiative activities, an event describing the occurrence of a Christmas Day parade in San Jose would not be relevant even if the actors, action, location, time, and all other fields were described accurately. This note on relevance, though introduced in a discussion on IIE systems, could apply to all IE systems.

There are also additional metrics for user-oriented learning-based IIEEs. Pierce and Cardie conceptualize the efficacy of a system in two parts: coverage and responsiveness (Pierce and Cardie, 2001). Coverage is “the system’s ability to extract all desired information for the user, i.e. to completely cover the task” (Pierce and Cardie, 2001, 1). Coverage could be measured by precision, recall, or F1 Score; it is describing accuracy. Responsiveness is “the system’s ability to achieve a reasonable level of performance without undue burden upon the user” (Pierce and Cardie, 2001, 1). When considered within the logic of user-oriented learning, coverage encourages more training examples from the user whereas responsiveness encourages less. Culotta et al. describes a similar concept through corrective feedback, as the ability to solicit corrections from the user, and persistent learning, as the ability of the system to continually update its prediction model(s) (Culotta et al., 2006). Corrective feedback is dependent on the usability and utility of the IE system, as they determine the quantity and

quality of output data which will in turn become training data for the next iteration of the prediction model (Culotta et al., 2006).

So far, the discussion of metrics has been limited to the assessment of one IE task at a time. This is sufficient for a joint classification approach, but not a pipeline classification approach in which multiple steps exist. In pipeline classification, we must understand how the performance of an upstream module influences – for better or worse – the performance of a downstream module, and how any changes to upstream modules may affect downstream modules (Xiang and Wang, 2019, 11; Liu, Luo, and Huang, 2018; Du and Cardie, 2020). Ahn presents a method for evaluating the effect of changing individual modules on overall performance as measured by ACE value in a traditional automated IE system (Ahn, 2006). However, this procedure for a default pipeline classification, user-oriented learning-based IIEE remains as a gap in the literature compared to the much heavier documented topics of joint classification and fully automated IE systems. The following sections of this paper attempt to address this gap through a discussion of the development of the SCOPE tool and the implementation of metrics for coverage and responsiveness to assess the efficacy within and across the tools’ modules.

## 2 Methods

The core goal of this paper is to implement a test of the efficacy of the SCOPE interactive information extraction environment (IIEE), relative to fully manual workflows. The results of this test will demonstrate the ability of SCOPE to present the relative strengths and weaknesses of manual users and AI through the establishment of baselines for future comparison, as well as provide precedent and methods for integrating the two. In this section, I first introduce the broad design strategy for SCOPE, as well as the tool’s existing and proposed infrastructure, insofar as it is necessary to enable a discussion of the findings of my analysis. Second, I outline the approach used to assess the efficacy of the IE methods implemented within the SCOPE tool, and how this approach could be extended to future AI improvements to IIEEs.

## 2.1 Introducing the SCOPE Method

In this brief section, I provide a broad overview of the design decisions of SCOPE as they pertain to the efficacy assessment methods described later in Section 2.2. A full discussion of the design rationales and decisions for SCOPE is provided in Appendix B. A complete visual documentation of the SCOPE tool and its parts is provided in Appendix C. All of the code is accessible at <https://github.com/wmgeolab/scope>.

As a user-oriented learning-based IIEE, the efficacy of the SCOPE tool is a function of its ability to maximize coverage and responsiveness (Pierce and Cardie, 2001). This ability starts with eliminating the need for the creation of a corpus of hundreds of annotated training documents by instead learning as the user manually completes their work. The machine learns through consistent interaction with the user in the form of training data and quality assurance (Culotta et al., 2006). Figure 1.3 below illustrates how the SCOPE tool works to maximize coverage and responsiveness in each of its modules. Users operate the tool to manually complete the information extraction task and the output of their work is used as the training data for machine-assisted information extraction. A sample of the data (denoted by the dotted lines) from both the manual and automated workflows is quality assured by manual users before being added to the training data. A deeper explanation is provided in Appendix B.

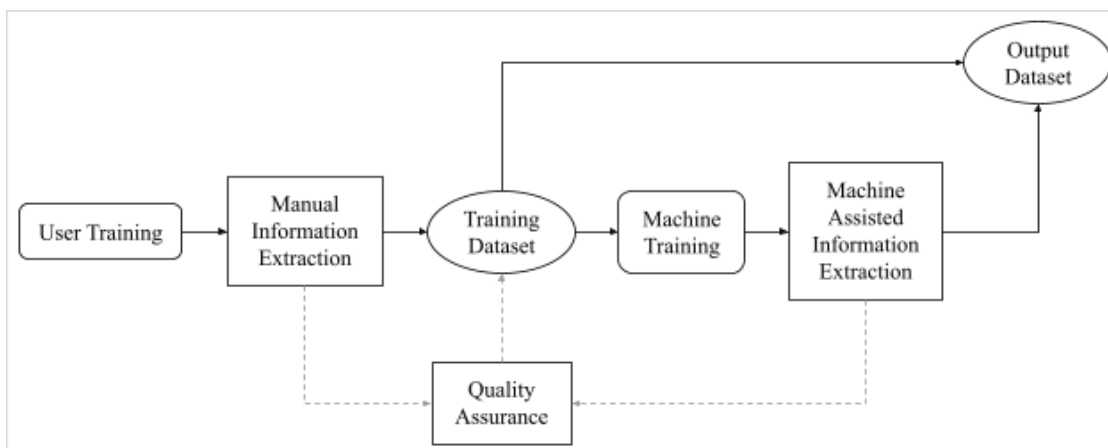


FIGURE 1.3: SCOPE's User-oriented Learning-based System

The SCOPE tool itself is developed in Python's Django web framework, which

uses a system of models, views, and templates to define how data is stored, processed, and interacted with in various parts of the tool. To maximize flexibility while benefiting from user-oriented learning, the SCOPE tool is developed in independent, optional “modules” which are linked together by their expected inputs and outputs. Each module is a grouping of functionalities which enable users to process or interact with the data in some specific, predetermined manner. All modules also belong to one of four “classes”: Administration, Pipelines, Workflows, or Analytics. Administration modules enable a user to set up all the basic infrastructure of their version of the SCOPE tool; connecting to a database, adding user credentials, selecting modules to include, and so on. Pipeline modules bring in massive amounts of online source data into the database. Workflow modules cover the entirety of the traditional IE task; importing sources from the database into the tool’s manual and automated workflows, and extracting and parsing information from those sources. Analytics modules offer additional tasks, such as geocoding event locations or analyzing sentiment in text. A fuller discussion of each module is provided in Appendices B and C, including a list of which modules have already been constructed and which are proposed to be developed next.

Across its modules, the SCOPE tool follows by default a pipeline classification approach to information extraction, as opposed to a joint classification approach. In a pipeline classification approach, each classifier - which we call modules - is designed to accomplish a specific subtask (Xiang and Wang, 2019; Riedel and McCallum, 2011). For example, one module identifies relevant information and extracts it from source documents (Extracting) while another codes that information into a structured data observation (Parsing). As detailed in Appendix B, following a pipeline approach enables the SCOPE tool to be tailored to user-oriented learning situations which do not require the full EE process of transforming unstructured source text into structured events data, whereas a joint classification approach would be limited here as it handles the entire EE process simultaneously. In other words, pipeline classification enables users to decide for themselves which modules to include in their version of the SCOPE tool. A user-oriented learning approach also overcomes the heavy feature engineering and linguistic knowledge usually required in each stage of a pipeline approach (Cardie and Pierce, 1998; Du and Cardie, 2020).



It is relevant to note that there is substantial literature which indicates that a pipeline approach to information extraction can be particularly susceptible to error propagation, “where errors in an upstream classifier are easily propagated to those downstream classifiers and could degrade their performance” (Xiang and Wang, 2019, 11; Liu, Luo, and Huang, 2018; Du and Cardie, 2020). A separate set of literature indicates that this weakness can be overcome, and there are at least two solutions to this problem to maintain SCOPE’s coverage without sacrificing responsiveness (Riedel and McCallum, 2011; Venugopal et al., 2014). First, the inclusion of modules for quality assuring the outputs of each suite (e.g., Extracting (QA) Module, Parsing (QA) Module) can mitigate the effects of error propagation by subjecting each step of the information extraction process to one or more rounds of quality assurance by multiple manual or automated coders. Second, it is within the capabilities of the SCOPE tool to simply create a single suite of modules which contains the entire IE system (a joint classification approach). In this suite, one module would be dedicated to manually parsing structured events data from unstructured source text. Another module would house the machine learning model which automatically completes this same process. A third module would offer the ability to quality assure the outputs of the suite.

Another potential weakness of the pipeline approach might be that a downstream module cannot impact upstream modules’ decisions, and the various interdependencies of different IE subtasks cannot be well utilized (Ahn, 2006; Xiang and Wang, 2019; Venugopal et al., 2014). This is a recognized limitation of the SCOPE tool as it is presently designed and is accepted by other scholars as well (Ahn, 2006; Aggeri et al., 2016; Riedel and McCallum, 2011). Currently, the best method for ensuring efficacy of all modules is to maximize the efficacy of upstream modules, including the acquisition and storage of any data which may be valuable downstream. The coverage and responsiveness of upstream modules indirectly impacts those downstream, so optimizing the methods – manual and/or automated – of each module is critical.

## 2.2 Proceduralizing the Assessment of the Efficacy of IE Methods in SCOPE

In this section, I build on existing literature to propose a design for assessing the efficacy of the methods applied in the SCOPE tool. These methods can be applied to a fully manual, automated, or hybrid SCOPE workflow.

To begin, I separate efficacy into two systematized concepts: coverage and responsiveness (Pierce and Cardie, 2001). Coverage describes the tool's accuracy, or ability to accomplish the IE task. Responsiveness describes the tool's ability to achieve a certain level of accuracy given a set schema and interaction with users. Next, I deconstruct each of these concepts into operational measures. Following precedent in IE literature, I measure accuracy by precision, recall, and F1 Score (Ahn, 2006; Campos et al., 2014; Culotta et al., 2006; Hogenboom et al., 2016; Pafilis et al., 2016). I use time as the measure for interaction with users (Campos et al., 2014; Culotta et al., 2006; Hogenboom et al., 2016; Pafilis et al., 2016).

In a pipeline classification approach, efficacy can be measured for the entire IE pipeline as well as for individual modules (Xiang and Wang, 2019). The efficacy of each module will impact the overall efficacy of the pipeline, with upstream modules being especially influential on downstream modules (Xiang and Wang, 2019, 11; Liu, Luo, and Huang, 2018; Du and Cardie, 2020; Riedel and McCallum, 2011; Ahn, 2006). As such, a change in the efficacy of one module can change the efficacy of subsequent modules and the entire pipeline. Further, the methods used for measuring precision, recall, and F1 Score are dependent on the unit being examined (i.e., a specific module or the entire pipeline). For example, assessing an Extracting module requires determining whether all the relevant information describing an event and no irrelevant information is included within each extract. On the other hand, assessing a Parsing module requires determining whether each field (e.g., event type, actors, location, date) was properly classified based on the information in the input extract data. Of course, assessing a joint classification approach would require determining whether each event was properly identified, extracted, and parsed in one step from the source text.

For the purposes of this paper, I focus on assessing the SCOPE tool's Extracting module. While the methods for measuring accuracy are unique to the

Extracting module, the overall process can be applied to any module or the entire workflow of the SCOPE tool. I choose to focus on the Extracting module as it is one which we could establish a baseline for both the manual and prototype automated methods.

To measure the accuracy of the Extracting module, I begin by setting criteria for an event and an extract. An event is an occurrence of something at a certain time and place involving one or more participants (Xiang and Wang, 2019). An extract is an unstructured text string composed of unaltered sentences extracted from the source which should include information on one event, or multiple events when they are listed together and difficult to separate without losing information (Liu, Luo, and Huang, 2018). Accordingly, precision is the fraction of extracts which properly describe relevant events and recall is the fraction of properly described relevant events that are retrieved (Hogenboom et al., 2016). Figure 1.4 presents a confusion matrix for the Extracting modules.

		<u>Actual</u>	
		Positive	Negative
<u>Predicted</u>	Positive	Event information properly extracted	Event information extracted which should not have been
	Negative	Event information not extracted which should have been	Irrelevant information not extracted

FIGURE 1.4: Confusion Matrix for the Extracting Module

To calculate these requires assessing the event information which should be extracted as well as the events and other information which should not be extracted. In other words, evaluating a *properly extracted* event from an improperly extracted event. For the purposes of this paper, I set the rules that – after removing stop words, repeated words, and punctuation – at least 67% of the extract must be necessary text describing the event, no more than 20% can be text which is irrelevant to the event, and no more than 33% of the necessary text can be missing from the extract. Sources with only irrelevant information should have no extracts. This method loosely follows that of Jaccard similarity and cosine similarity in information retrieval and text mining literature, where

the similarity between the contents of documents is calculated (Chahal, 2014; Gomaa and Fahmy, 2013). The method put forth in this paper is stricter than Jaccard similarity because it differentiates between appropriate, inappropriate, and missing text, whereas Jaccard similarity is only a measure of overlap and does not assess text differently. I argue that this method is appropriate because the inclusion/exclusion of appropriate/inappropriate information can impose different effects on the IE process.

A challenge to this approach is that different human and automated coders will likely not create the same number of extracts for a source, and their extracts may describe different combinations of events. The order of the extracts could be identified and corrected through calculating the similarity between each combination of extracts, but – to the best of the author’s knowledge – there is no documented method of accurately assessing which extracts are describing which events if the number of extracts differs between the two datasets. This may not be an issue, though, given that some extracts may already contain multiple events and some events may possess the same descriptive information and interdependencies. Further, event information could still be parsed correctly if a “Deduplicating” module is included prior to the Parsing module to merge together extracts which describe the same event(s) but were incorrectly extracted separately. As such, I propose to evaluate the content of the extracts together (as opposed to individually) to mitigate this challenge. Whether a penalty should be imposed for when the number of extracts is different from the key (i.e., validation data) should be a focus of future research, especially with regard to its potential impact on the Parsing module.

Now that we have a standardized method for measuring the accuracy of a module, we are able to contrast this with the time it takes to complete the IE task to determine the efficacy of the module. The first run of the manual version of the module (or fully manual workflow) is the baseline for comparison. When changes to the SCOPE tool are made, any resulting increase in accuracy or decrease in completion time constitutes an improvement of the tool’s efficacy. This includes user interface changes, development of machine learning models, or even improvements to the processing capabilities of the database infrastructure.

## 2.3 Case Example: Developing a Comparative Efficacy Metric for Manual and Machine-Automated Event Extraction

In the remainder of this paper, we implement the SCOPE tool with a first run of its manual workflow to establish a baseline for future comparison. We construct a prototype function to illustrate a comparison between human and automated methods. It is important to note that the majority of the tool is still in ongoing development, especially the automated components, and the version we present here represents the basic infrastructure of the tool. Nonetheless, establishing methods for how to assess and compare the results of the tool is a vital step to take as we begin to develop the more advanced automated components.

### 2.3.1 Establishing a Baseline Using SCOPE for Manual Event Extracting and Parsing

Upon completion of the initial version of the SCOPE tool presented in this paper, we set out to begin integrating the tool into the existing operations of the author's research team, geoParsing. Over the course of six weeks in March and early April 2021, the team transitioned from their previous spreadsheet-based manual IE system described in section 1.1.1 to using the SCOPE tool for extracting events on topics of Chinese Belt and Road Initiative development projects in Latin America and the Caribbean and Russian foreign relations in the Central African region. This transition took place at weekly team meetings.

For the first two weeks, the researchers were introduced to the tool by the author and practiced navigating its different pages and functionalities for thirty minutes each week. During this time, the researchers also provided constructive feedback about the tool's design, adjustments which would improve navigation and usability, and any error messages they encountered. For example, one researcher suggested that we add an asterisk (\*) next to the required fields in the Manual Parsing module. Another researcher was experiencing difficulties when formatting their event dates, so we included a calendar widget for easier use. Many researchers ran into error messages after forgetting to log into their GitHub account or attempting to checkout multiple tasks at the same time, so we implemented protections to protect against these errors.

For the remaining four weeks, the researchers used the SCOPE tool to extract and parse events from a prepared dataset of online news sources. The author provided a quick recap tutorial at the first of these four weekly meetings and was present each week to troubleshoot any potential errors. The expectation was for each of the 15 researchers to complete the manual workflow for 4 sources each week for a total of 240 sources. These sources were from a mix of news outlets and curated based on the domain of the researchers, with 112 on topics related to China and 128 on topics related to Russia. Some of the sources were purposefully irrelevant and it was the responsibility of the researchers to assess whether each source had relevant event information to extract and parse. Due to scheduling constraints of team members, we decided to hold time constant so each researcher spent 4 minutes on extracting and 4 minutes of parsing for a total of 8 minutes.

By integrating SCOPE into the existing manual methods, we were able to develop a reasonable standard for future comparisons. For the purposes of this paper, we continue to focus our assessment of efficacy on the Extracting module. Human researchers are not 100% accurate all the time when extracting information. One cannot expect a machine to be 100% accurate either. To develop the answer key, 30 sources were randomly selected to be extracted by the author. Rather than holding time constant again, the author spent as much time as was necessary to properly extract all relevant event information from each of the sources. The average of these times is the time required to achieve 100% in precision and recall. We then plot this point against the results of the time-constrained researchers to establish a baseline to compare future improvements against.

### **2.3.2 Constructing Prototype Machine-Automated Functions**

We also developed a prototype “auto-assist” function to demonstrate a comparison between human and automated methods. For illustrative purposes more than productivity, the function is incredibly simple. Given a predetermined list of trigger words, it scans each of the 30 sources, detects every time a trigger is present, and extracts the three sentences including and following the trigger’s

location in the source. Those three sentences constitute its extract.<sup>4</sup> We had no expectation that this function would outperform the 15 researchers with regard to accuracy, but it should possess a massive advantage in performance time. The results of the function are plotted against the manual results.

### 3 Results and Discussion

In the remainder of this paper, I provide the results of the two tests to establish baselines for manual and automated IE tasks. I discuss the meaning of these results, address remaining limitations, and conclude with remarks on future work. The focus is again on the overall process through which we are able to assess the efficacy of the SCOPE tool, rather than the actual performance of this largely illustrative exercise using the initial version of SCOPE.

Of the 240 sources meant to be processed by the 15 researchers, only 207 were processed. 12 of these were excluded because researchers had to be absent from a weekly meeting. The remaining 21 were not processed due to various inefficiencies (e.g., error messages) during the first two weeks of using the SCOPE tool on one of geoParsing’s subteams. Given scheduling constraints of the researchers, these sources could not be made up and were excluded from the testing. They are not represented in the results. Another 45 sources were excluded from the random analysis because their websites did not allow the SCOPE tool’s built-in import function to automatically scrape their source text. The prototype “auto-assist” function uses this scraped source text as an input. The subset of 30 sources was randomly selected from a pool of 162 sources.<sup>5</sup>

	Manual	Manual (SW)	Auto-Assist	Auto-Assist (SW)
Precision	0.269	0.269	0.034	0
Recall	0.388	0.318	0.077	0
F1-Score	0.318	0.292	0.048	0

TABLE 1.2: Accuracy Statistics for Each Method

<sup>4</sup>The function relies on the *requests* and *biolperpy3* packages to retrieve the source text. It uses the *find* and *sent\_tokenize* modules from the *nltk* package to extract the sentences.

<sup>5</sup>All data and code for this section can be found at [https://github.com/wmgeolab/scope/tree/master/resources/scope\\_testing\\_results\\_IMPORTANT](https://github.com/wmgeolab/scope/tree/master/resources/scope_testing_results_IMPORTANT).

Table 1.2 provides the precision, recall, and F1 Score attained by the researchers for the random subset of 30 sources. These statistics were achieved by grouping extracts from the same source together, tokenizing them, removing duplicate words, removing stop words (SW), and finally calculating the percentages of token counts which met each criteria. Overall, the precision was 26.9%, recall was 38.8%, and F1 Score was 31.8%; when stop words were removed, the results changed as such – 26.9%, 31.8%, and 29.2%. The answer key took an average of 3 minutes and 16 seconds per source to extract; the median time was 2 minutes and 54 seconds. The times for each source are shown in Figure A.7. Surprisingly, both the researchers and the key extracted a total of 88 extracts from the 30 randomly selected sources, though the number of extracts does not remain consistent for every source. The number of extracts by source can be seen in Table A.1. Clearly, the low performance of the researchers despite taking more time than the answer key required illustrates the need for improvement using the SCOPE tool to integrate more advanced machine learning-based methods.

As expected, the auto-assist function performed far below the manual researchers in terms of accuracy with a precision of 3.4%, recall of 7.7%, and F1 Score of 4.8%; these dropped to 0 when stop words were removed. It struggled especially with websites which loaded their content dynamically using javascript and websites which were originally in a foreign language, such as French. The function also could not distinguish between events which were irrelevant to the project and should not have been extracted. Both the manual researchers and the prototype function achieved markedly higher recall than precision within the subset of 30 sources, suggesting a tendency to extract more information from the source than was appropriate and thus increasing the number of events extracted at the expense of always extracting the appropriate amount of information. The 67-20-33 percentage rule is especially strict when grading the precision of the method, as 67% of the text must be appropriate and no more than 20% can be inappropriate. On the other hand, up to 33% of appropriate text could be missing from the extract when grading recall. The prototype function performed incredibly quickly. The average time was 3.25 seconds; the median time was 3.11 seconds. As it was not our intention to judge the prospect of auto-assisted event extraction by the performance of the basic function here, it should instead be viewed as a baseline and an example of how one could assess a more



sophisticated machine learning model in the future.

An improvement can be increased accuracy (i.e., precision, recall, or F1 Score) or decreased completion time. For example, the baseline F1 Score results for each method can be seen in Figure 1.5. Decreasing the amount of time required for manual researchers to achieve the same level of accuracy could be achieved through the implementation of a more robust version of the prototype function to detect event information from source text. Likewise, the level of accuracy achieved by the automated method could be improved by integrating the feedback of manual researchers, even if it results in a slightly slower completion time. Similar figures for each of the measures of accuracy can be seen in Appendix A. These results suggest that with some effort, an automated method could readily benefit EE tasks as training data is amassed for the development of advanced machine learning-based models and fully automated modules. These methods can be evaluated using the same measures presented here. The wide gaps between the key, the manually extractions, and the automated extractions illustrates the high potential value of the SCOPE tool going forward as its remaining components are developed.

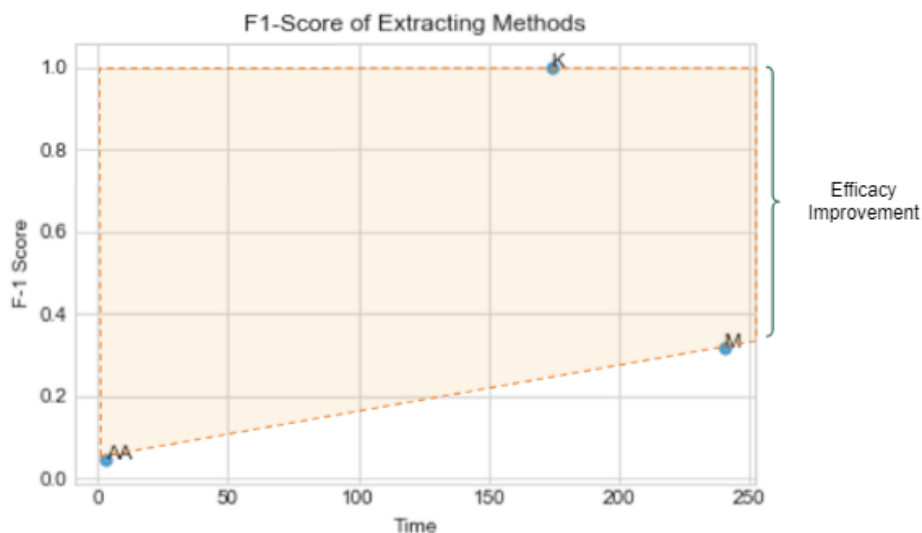


FIGURE 1.5: Plot of F1 Score and Speed Attained by Each Method

### 3.1 Conclusion

There are several areas of improvement for SCOPE in future work. These areas can be split into structural, methodological, and measurement improvements. Structurally, a limitation of the current SCOPE tool is that workflows are immutable once created. In this paper, we demonstrate how different functions and models can be improved. However, the tool is not yet able to accommodate changes to schemas or which modules are included. The next iteration of SCOPE should resolve this limitation by enabling users to retroactively make changes and have the data either be preserved or automatically adjusted. Further, not all the proposed modules have been developed. Future work should prioritize the development of remaining Administration modules and early-stage Auto Workflow modules, as they collectively comprise the most necessary functionalities for the event extraction task.

Methodologically, there is clear room for improvement with regard to the machine learning models. The prototype auto-assist function applied here was for the demonstrative purpose of establishing a baseline for comparison rather than achieving practical and accurate results. The development of superior machine learning models was outside of the scope of this thesis to accomplish, but they are integral to the success of SCOPE, particularly as a user-oriented learning-based IIEE. Future work should prioritize the development of these models – to be situated within the Auto Workflow modules – and study their efficacy over time as quality assured data accumulated through users' interactions with the SCOPE tool is iteratively introduced as training data.

With regard to measurement improvements, currently we can only assess the predicted change in efficacy of a module or workflow following an improvement, but mapping the exact way a module-level change impacts succeeding modules or the entire workflow is more ambiguous. Further exploration of these intermodular interactions would make a meaningful contribution to the literature. Lastly, future work might consider integrating additional measures for interaction with users beyond time. These measures could strengthen our conceptualization of responsiveness by integrating structured insights of how user-friendly and operable the SCOPE tool is whenever a user-facing update is made and by tracking how many observations are required to reach a certain level of accuracy based on user-oriented learning. Any update which makes the

tool more user-friendly or less observation-dependent will improve the tool's responsiveness.

We set out in this paper to present the ongoing development of a novel user-oriented learning-based IIEE which encompasses all the tasks involved in operating an IE project and integrates manual and automated IE methods to optimize efficacy. SCOPE, a Python Django-based tool divided across specialized modules accomplishes this goal. Though only the manual workflow is operational thus far, SCOPE poses to have high potential in the field of IE, specifically for the creation of events datasets. To the best of the author's knowledge, SCOPE is uniquely situated among IE technologies as a system which is flexibly built to enable manual, automated, and/or interactive workflows while also incorporating a user-oriented learning approach to iteratively improving the tool's automated components over time. These capabilities are made possible by the tool's modular design, in which modules are developed independently and their inclusion in the tool can be tailored to specific project needs. We demonstrate the SCOPE tool's potential through the proceduralization of assessing efficacy of its methods, whether applied at the module or workflow level. In doing so, we lay the foundation for integrating and testing the results of future improvements to the tool, whether they be updates to the user interface, advancements in the machine learning models, or any other changes. We then put this proceduralization into practice by establishing baselines for future comparison for the manual and automated versions of the Extracting modules, achieving 31.8% and 4.8% F1 Score accuracy within 240 seconds and 3.25 seconds respectively. These results highlight the comparative advantages of both methods and suggest the value of an integrated approach to IE. Lastly, the proposed inclusion of Analytics modules to handle additional tasks such as geocoding event locations, identifying misinformation, or analyzing sentiment in text situates the SCOPE tool to in the future be able to accomplish many innovative tasks beyond the central goal of extracting events information from text.

## 4 Acknowledgements

I would like to thank the committee (Dan Runfola (chair), Anthony Stefanidis, and Maurits van der Veen), Karim Baghat, the geoParsing Team @ geoLab, the Roy R. Charles Center for Academic Excellence, the faculty of the Department of International Relations and the Data Science Program at William & Mary, select faculty and researchers at William & Mary's Global Research Institute, friends, and family for their support of this project and preceding work.

Dan Runfola, Assistant Professor of Applied Sciences and faculty director of William & Mary's Geospatial Evaluation and Observation Lab, has been an invaluable mentor and supporter throughout my undergraduate career. His impact has been at least three-fold. First, as the advisor to my self-designed Bachelor of Arts in Data Science, guiding me to a balanced and enriching undergraduate curriculum. Second, as the faculty director of geoLab, entrusting me to found and lead the geoParsing Team. Third, as the advisor to SCOPE, entrusting me to found and lead yet another ambitious project. In each of these roles, he has offered his time, resources, and genuine care.

Karim Baghat, a PhD candidate working with Dan Runfola, has been an incredible mentor throughout this project. He was especially valuable to the design and development of the SCOPE Django website, as well as a brainstorming partner in the early stages of the project.

The geoParsing Team in William & Mary's Geospatial Evaluation & Observation Lab (geoLab) was both the motivation and a vital component for this project to be completed. Members of the INNOVATION Program (Landon Clime (lead), Moses Alexander, Monica Alicea, and Gracie Kosco) supported the development of the Scraping and Sourcing Modules. Members of the BRIGHT Project (Caroline Morin (lead), Remy Fritz (assistant lead), William Weston, Asha Silva, Wendy (Yiwen) Sun, Kaitlyn Wilson, and Sophie Pittaluga) assisted in the manual testing of the SCOPE tool, specifically for China-focused sources. Members of the TRACAR Project (Cole Spiller (lead), Garrison Goetsch (assistant lead), Monica Sandu, Erin Horrigan, Amelia Grossman, Yasha Barth, Zoe Roberts, and Aliia Woodworth) assisted in the manual testing of the SCOPE tool, specifically for Russia-focused sources. Alumni of the geoParsing Team (including Kate Munkacsy, Celia Metzger, Greyson Pettus, Emily Maison, and Olivia Hettinger)

were also instrumental to earlier work which preceded this thesis. Other alumni of the geoLab (including Rachel Oberman, Joshua Panganiban, Heather Baier, John Napoli, and Lauren Hobbs) were also sources of support to earlier work which preceded this thesis.

Earlier work which preceded this thesis was partially funded by the Charles Center at William & Mary through the Monroe Scholars Program. Beyond the scope of this thesis, faculty of the Charles Center (including Lindsey Love, Christine Azznara, Dan Cristol, and Chelsea Craddock) were meaningful sources of support to my undergraduate career.

Beyond the scope of this thesis, I also thank numerous professors, faculty, and staff at William & Mary, many of whom belong to the Department of International Relations and Data Science Program. Here, I highlight a few. Mike Tierney, Professor of International Relations and Director of the Global Research Institute (GRI), has been an invaluable mentor and supporter throughout my undergraduate career. As the advisor to my Bachelor of Arts in International Relations, he has time and time again offered his time, resources, and genuine care. Also at the GRI, David Trichler, Rebecca Halleran Latourell, and Carrie Dolan have been great supporters and mentors as well. Tyler Frazier, Dennis Smith, Eric Arias, Chinua Thelwell, Francis Tanglao Aguas, and Philip Roessler stand out among my most supportive instructors.

Last, but certainly not least, I would like to thank friends and family. My brother, my first and continuing role model. My mother, my inspiration and number one supporter.

# Appendix A

## Main Appendix

The screenshot shows a Google Sheet interface with the following data:

Source ID	Source URL	Publication Date	Relevance	Asides (only if no activities)	Source Name	Source PDF	Source Type	Corresponding Activities	Additional Notes
273	https://barbadostoday.bb/	8/11/2019	Yes		Barbados Today	https://drive.google.com/	4. Media reports		96
274	https://www.gaia.bb/content	N/A	Yes		GAIA inc.	https://drive.google.com/	3. Implementing or		97
275	http://biba.bb/mora-chine	11/23/2015	Yes		Barbados International Business Association (BIBA)	https://drive.google.com/	7. NGO or civil soci		97
276	https://www.stlucianews.com	2/14/2019	Yes		St Lucia News	https://drive.google.com/	4. Media reports		97
277	http://www.chinadaily.com	3/25/2016	Yes		China Daily	https://drive.google.com/	4. Media reports		97
278	https://wlcnews.com/can	9/6/2019	Yes		WLC News	https://drive.google.com/	4. Media reports		98
279	https://www.dominicavibe.com	8/9/2016	Yes		Da Vibes - The Caribbean's News Portal	https://drive.google.com/	4. Media reports		98
280	http://www.jamaicaobserver.com	8/9/2016	Yes		Jamaica Observer	https://drive.google.com/	4. Media reports		98
281	https://www.dominicavibe.com	9/13/2016	Yes		Da Vibes - The Caribbean's News Portal	https://drive.google.com/	4. Media reports		98
282	http://en.hrbti.com/news/	9/7/2018	Yes		Hiernan Provincial Communications Planning, Survey, and Design Institute	https://drive.google.com/	3. Implementing or		99
283	https://www.dominicavibe.com	6/13/2018	Yes		Da Vibes - The Caribbean's News Portal	https://drive.google.com/	4. Media reports		99
284	https://sundominica.com/ai	7/24/2018			The Sun	https://drive.google.com/	4. Media reports		99
285	http://news.gov.dm/news/	11/25/2016			Government of the Commonwealth of Dominica	https://drive.google.com/	2. Official from Chir		100
286	https://www.caribjournal.com	6/16/2014	Yes		Caribbean Journal		4. Media reports		101
287	https://www.caribjournal.com	05/05/2015	Yes		Caribbean Journal		4. Media reports		101
288	https://www.scmp.com/or	6/18/2014	Yes		South China Morning Post		4. Media reports		101
289	https://www.dominicavibe.com	11/18/2013	Yes		Dominica Vibes		4. Media reports		102
290	http://www.jamaicaobserver.com	08/26/2014	Yes		Jamaica Observer		4. Media reports		102
291	https://china.aiddata.org/	2017	Yes		AidData		5. Peer-reviewed sc		102
292	https://www.dominicavibe.com	1/20/2016	Yes		Da Vibes - The Caribbean's News Portal	https://drive.google.com/	4. Media reports		103
293	https://pointville.ag/2018/		Yes		Pointe Express		4. Media reports		104

FIGURE A.1: Source Worksheet of geoParsing’s BRIGHT Google Sheet

Activity ID	Source ID	Number of Sources	Activity Name	Funder Type	Funder(s)	Implementer(s)	IRI Status	Activity Type	Financial Type	Dollar Amounts	Country Code	Country	Specific Location	Notes	Activity Date 1	Activity Date 2	Activity Status
1	111561567	1	Fuente Amador Cruise Terminal	Other Chinese Institution	China Harbour Engineering Company	China Harbour Engineering Company	1.4	Port, Tourism	Contract	\$165,000,000 USD	PAN	Panama	Fuente Amador Cruise Terminal, Amador, Panama City, Panama	China Harbour Engineering Company (which help from Beijing company, see De Nullo) is building a port for cruise ships, which would also convert Perico Island into a tourist destination. Organized directly by meetings between the Panamanian and Chinese government. This port could accommodate 2 mega cruise ships and 10000 passengers, with the eventual goal of a five mega cruise ship capacity in the future. Following Chinese President Xi Jinping's first visit to Panama, the government of Panama awarded a contract to a Chinese consortium to build the fourth bridge over the Panama Canal. The 6.5 km bridge, built just north of the Bridge of the Americas, will include six lanes of traffic and an extension of the Panama Metro. The bridge will connect Panama City and its western suburbs.	10/16/2017	N/A	Under construction - delayed
4	211601169 31791365	2	Panama City Bridge	Other Institution	* Govt of Panama	China Harbour Engineering Company	1.3	Bridge	Contract	\$1,400,000,000 USD	PAN	Panama	Panama City Bridge, Panama City, Panama	The planning and construction of the Amador Convention Center located along the Amador Causeway, a four-mile-long thoroughfare that bypasses the Panama Canal. From the Pacific Ocean and contains a concentration of Panama City's tourism attractions, began in 2011. However, shortly after, the project was abandoned due to liquidity issues. In 2016, the contract was picked up by Chinese companies, with an expected finish date in 2018. Due to issues with the structural integrity, the grand opening was delayed until September of 2019 and now the convention center is completed and open. Initially the project cost about \$193.7 million, but costs grew to about \$215 million after delays and other issues. Panama signed an explicit MOU with PRC the year after this project started and CCCCC is a state-owned company.	12/03/2018	N/A	Cancelled
6	31801841 31791365	2	Trans-Panama High-Speed Railway	Other Chinese Institution	* PRC	China Railway Design Corporation	1.1	Rail	Vague	\$4,500,000,000 USD	PAN	Panama	Panama City, Panama to David, Panama	China Railway Design Corporation's \$16 million feasibility study for a Trans-Panama high-speed railway was recently approved by the Panamanian government. The railway will stretch 391 km between Panama and Costa Rica. ORC is a state-owned subsidiary of CRCC.	03/27/19	N/A	Cancelled
7	414171172	1	Amador Convention Center	Other Institution	* Govt of Panama	China Construction America (affiliated to China State Construction Engineering Corporation (CCCC))	1.3	Recreational	Contract	\$208,700,000 USD	PAN	Panama	Amador Convention Center, Panama City, Panama	The planning and construction of the Amador Convention Center located along the Amador Causeway, a four-mile-long thoroughfare that bypasses the Panama Canal. From the Pacific Ocean and contains a concentration of Panama City's tourism attractions, began in 2011. However, shortly after, the project was abandoned due to liquidity issues. In 2016, the contract was picked up by Chinese companies, with an expected finish date in 2018. Due to issues with the structural integrity, the grand opening was delayed until September of 2019 and now the convention center is completed and open. Initially the project cost about \$193.7 million, but costs grew to about \$215 million after delays and other issues. Panama signed an explicit MOU with PRC the year after this project started and CCCCC is a state-owned company.	05/01/16	06/30/2019	Completed
8	51517	1	SOVY Chiriqui Bridge Phase 2 (H Project)	Other Institution	* N/A	China Construction America (affiliated to China State Construction Engineering Corporation (CCCC))	1.3	Energy, general	Contract	\$200,000,000 USD	PAN	Panama	Chiriqui Grande Substation in Barro Colorado, Panama	Construction of a new 500/230 KV Chiriqui Grande substation, built as an extension of the existing 500/230 KV Panama III substation, as well as the installation of a State-Vol Contractor of 101 - 30 MW&M of Panama III substation have been delayed in response to Chinese Gov's failure to meet the minimum requirements of the tender. While pending stages of February of 2018, the current tenders are being reviewed as of September 2018. CCT is a subsidiary of the State Grid Corporation of China (SGCC), a PRC owned company.	02/03/2016	N/A	Delayed
9	6161173	1	Datin Container 2 Port	Other Chinese Institution	Shanghai	China Landbridge Overseas Commodities Construction Company	1.2	Port	Contract	\$1,000,000,000 USD	PAN	Panama	Panama Colon Container Port, Colon, Panama	The planning and construction of the Amador Causeway, a four-mile-long thoroughfare that bypasses the Panama Canal. From the Pacific Ocean and contains a concentration of Panama City's tourism attractions, began in 2011. However, shortly after, the project was abandoned due to liquidity issues. In 2016, the contract was picked up by Chinese companies, with an expected finish date in 2018. Due to issues with the structural integrity, the grand opening was delayed until September of 2019 and now the convention center is completed and open. Initially the project cost about \$193.7 million, but costs grew to about \$215 million after delays and other issues. Panama signed an explicit MOU with PRC the year after this project started and CCCCC is a state-owned company.	05/03/2016	10/31/2019	Under construction - delayed
9	7161174	1	Maitani Gas 2 Power Plant	Other Chinese Institution	Shanghai	Shanghai Electric Power Co. Ltd.	9	Energy, general	Contract	\$600,000,000 USD	PAN	Panama	Maitani Gas Plant, Panama	A new gas combined cycle power plant (CCGT) is being constructed near the Maitani Gas Plant. The project is being developed by the Chinese group, Shanghai Electric.	12/12/2016	N/A	Under construction - delayed

FIGURE A.2: Activities Worksheet of geoParsing's BRIGHT Google Sheet

Activity ID	Source ID	Activity Name	Country Code	Specific Location	Notes	Activity Status	Location Type	Geocoded Location	Geocoded Location Type	Geocoder Notes	Deliverable Mapping	ADM1	ADM1 (if applicable)	Completion Status
1	111561567	Fuente Amador Cruise Terminal	PAN	Fuente Amador Cruise Terminal, Amador, Panama City, Panama	China Harbour Engineering Company (which help from Beijing company, see De Nullo) is building a port for cruise ships, which would also convert Perico Island into a tourist destination. Organized directly by meetings between the Panamanian and Chinese government. This port could accommodate 2 mega cruise ships and 10000 passengers, with the eventual goal of a five mega cruise ship capacity in the future. Following Chinese President Xi Jinping's first visit to Panama, the government of Panama awarded a contract to a Chinese consortium to build the fourth bridge over the Panama Canal. The 6.5 km bridge, built just north of the Bridge of the Americas, will include six lanes of traffic and an extension of the Panama Metro. The bridge will connect Panama City and its western suburbs.	Under construction - delayed	PORT	<a href="https://satellite.com/panama/panama-city/11561567">https://satellite.com/panama/panama-city/11561567</a>	EOB	Used satellite imagery and followed google maps	Yes (Both)	Provincia de Panama	NA	Quality Assured by SPM
4	211601169 31791365	Panama City Bridge	PAN	Panama City Bridge, Panama City, Panama	The planning and construction of the Amador Causeway, a four-mile-long thoroughfare that bypasses the Panama Canal. From the Pacific Ocean and contains a concentration of Panama City's tourism attractions, began in 2011. However, shortly after, the project was abandoned due to liquidity issues. In 2016, the contract was picked up by Chinese companies, with an expected finish date in 2018. Due to issues with the structural integrity, the grand opening was delayed until September of 2019 and now the convention center is completed and open. Initially the project cost about \$193.7 million, but costs grew to about \$215 million after delays and other issues. Panama signed an explicit MOU with PRC the year after this project started and CCCCC is a state-owned company.	Cancelled	BOO	<a href="https://satellite.com/panama/panama-city/211601169">https://satellite.com/panama/panama-city/211601169</a>	EOB	Followed map of bridge here: <a href="https://www.google.com/maps/@9.134541,-79.514722,15z">https://www.google.com/maps/@9.134541,-79.514722,15z</a>	Yes (Geocoded)	NA	Quality Assured by SPM	
6	31801841 31791365	Trans-Panama High-Speed Railway	PAN	Panama City, Panama to David, Panama	China Railway Design Corporation's \$16 million feasibility study for a Trans-Panama high-speed railway was recently approved by the Panamanian government. The railway will stretch 391 km between Panama and Costa Rica. ORC is a state-owned subsidiary of CRCC.	Cancelled	RR	PAN_ADM0_2_0_0_0	PLU	Coded up to country level	No	NA	Quality Assured by SPM	

FIGURE A.3: Geocoding Worksheet of geoParsing's BRIGHT Google Sheet





Activity ID	Source ID	Activity Name	Country Code	Activity Location	Notes	Location Type	Geocoded Location	Geocoded Location Type	Geocoder Notes	Deliverable Mapping	ADM1	Completion Status
1	1401720205	Lengo Songo 88 SFM Radio Station	CAF	CAR	Radio Lengo Songo: Wiggister funded a radio station to support President Touadera in the elections. Russia also has a history of media control. Says it is to improve communication in CAR.	STNR	CAF-ADM0-1580548715-B1	PCLI	Geoboundaries v.3.0.0 was used. Coded up to country because there was no information on where the station is based.	No	NA	Quality Assured by SPM
2	717220	Miss Bangui Beauty Pageant	CAF	Bangui Stadium	This pageant was funded also by Lohaye himself to sponsor President Touadera on his election campaign. Many important Russian officials (security advisor, Calhoun) also attended. To symbolize CAR's new national direction and show Russia's support.	BLDG	<a href="https://api.github.com/repos/geoapiv3/geoapiv3/commits/549c093f">https://api.github.com/repos/geoapiv3/geoapiv3/commits/549c093f</a>	BLOB	GeoJSON was used based on information from Google Maps. Coded up to the city block because uncertain if all activities took place in the stadium.	Yes (Both)	Bangui	Quality Assured by SPM
3	34174	Mobile Clinics	CAF	Ouadda	In 2017/2018 a Russian humanitarian convoy went across CAR as mobile clinics carrying materials for the construction of hospitals and providing free care to residents of Ouadda. 20 instructors were stationed here to protect the donated hospital. To provide assistance to residents in need.	PPL	<a href="https://api.github.com/repos/geoapiv3/geoapiv3/commits/83346c359d0">https://api.github.com/repos/geoapiv3/geoapiv3/commits/83346c359d0</a>	PPL	GeoJSON was used based on information from Google Maps.	Yes (Both)	Haut-Kotto	Quality Assured by SPM
4	43174	Mobile Clinics	CAF	Brao	In 2017/2018 a Russian humanitarian convoy went across CAR as mobile clinics carrying materials for the construction of hospitals and providing free care to residents of Brao. To provide assistance to residents in need.	PPL	<a href="https://api.github.com/repos/geoapiv3/geoapiv3/commits/5d8536c1">https://api.github.com/repos/geoapiv3/geoapiv3/commits/5d8536c1</a>	PPL	GeoJSON was used based on information from Google Maps.	Yes (Both)	Vakaga	Quality Assured by SPM
5					In 2017/2018 a Russian humanitarian convoy went across CAR as mobile clinics carrying materials for the construction of hospitals and providing free care to residents of Ndere. To provide assistance to residents in need.		<a href="https://api.github.com/repos/geoapiv3/geoapiv3/commits/2726412281183942228">https://api.github.com/repos/geoapiv3/geoapiv3/commits/2726412281183942228</a>		GeoJSON was used based on information from Google Maps.		Bamingui-Bango	

FIGURE A.6: Geocoding Worksheet of geoParsing’s TRACAR Google Sheet

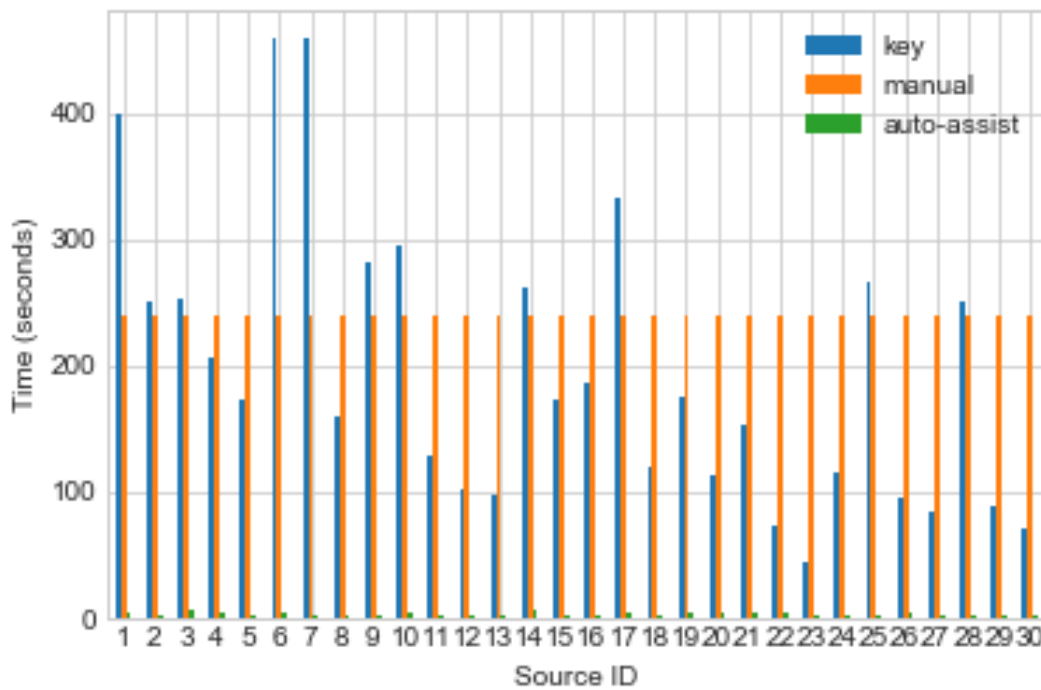


FIGURE A.7: Time Taken to Extract Information By Each Method

Test ID	Researchers	Key	Auto-Assist
1	3	2	4
2	2	2	10
3	3	1	4
4	3	1	5
5	4	1	3
6	2	3	16
7	7	10	1
8	5	1	1
9	2	5	6
10	8	8	10
11	4	1	12
12	3	4	14
13	2	2	10
14	3	6	7
15	2	3	8
16	1	2	1
17	4	6	23
18	5	1	8
19	1	1	11
20	3	2	9
21	2	4	8
22	0	0	14
23	1	0	17
24	1	1	2
25	1	2	4
26	1	1	0
27	4	5	23
28	3	5	8
29	1	1	8
30	1	1	8

TABLE A.1: Extract Counts by Source

	Manual	Manual (SW)	Auto-Assist	Auto-Assist (SW)
FP	19	19	28	29
FN	11	15	12	14
TP	7	7	1	0
TN	1	1	0	0

TABLE A.2: Accuracy Statistics in Terms of False Positives, False Negatives, True Positives, and True Negatives

	Manual	Manual (SW)	Auto-Assist	Auto-Assist (SW)
Precision	0.269	0.269	0.034	0
Recall	0.388	0.318	0.077	0
F1-Score	0.318	0.292	0.048	0

TABLE A.3: Accuracy Statistics for Each Method (repeat)

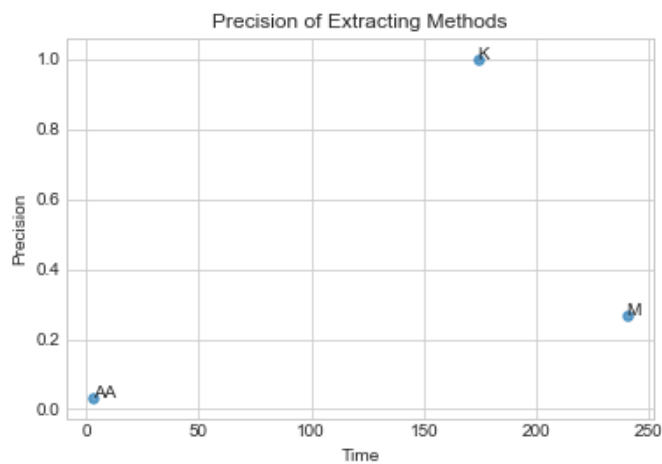


FIGURE A.8: Plot of Precision and Speed Attained by Each Method

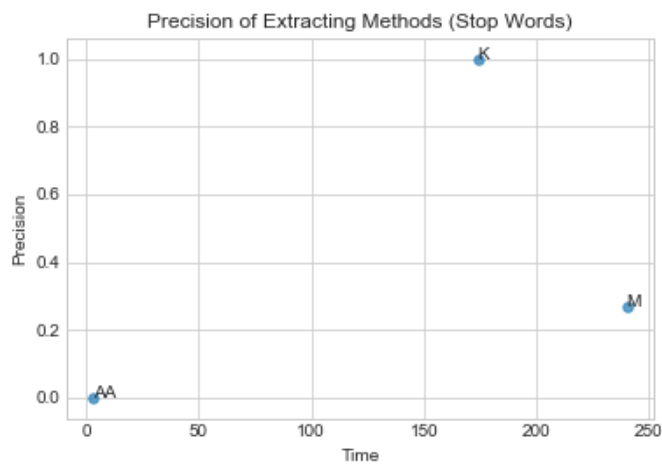


FIGURE A.9: Plot of Precision and Speed Attained by Each Method (Stop Words)

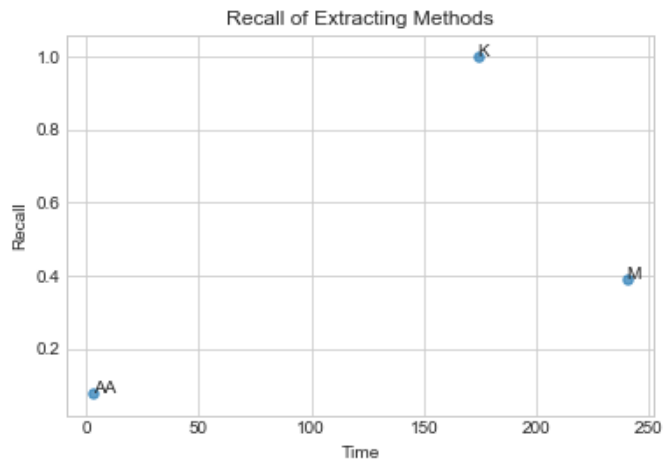


FIGURE A.10: Plot of Recall and Speed Attained by Each Method

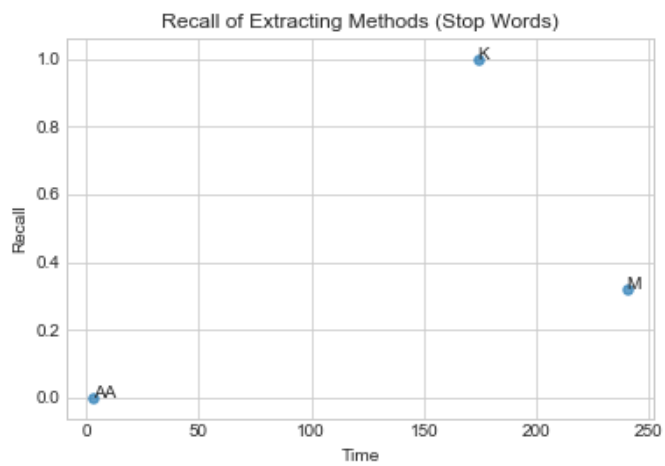


FIGURE A.11: Plot of Recall and Speed Attained by Each Method (Stop Words)

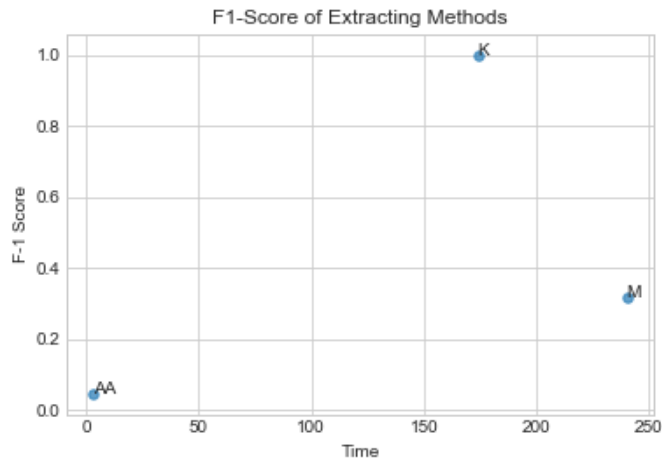


FIGURE A.12: Plot of F1-Score and Speed Attained by Each Method

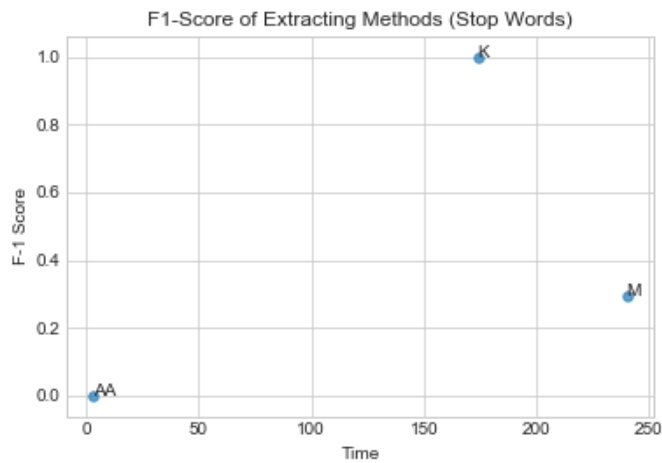


FIGURE A.13: Plot of F1-Score and Speed Attained by Each Method (Stop Words)

## Appendix B

# Designing the Scope Tool

### B.1 User-Oriented Learning-based System

The SCOPE tool was designed to provide project-tailored IE technology without need for expertise in NLP, computational linguistics, and data engineering. It follows the fundamentals of user-oriented learning, in which the machine learns through consistent interaction with the user in the form of training data and quality assurance (Pierce and Cardie, 2001; Culotta et al., 2006). Figure B.1 below illustrates the logic of the SCOPE tool, with each part described below.

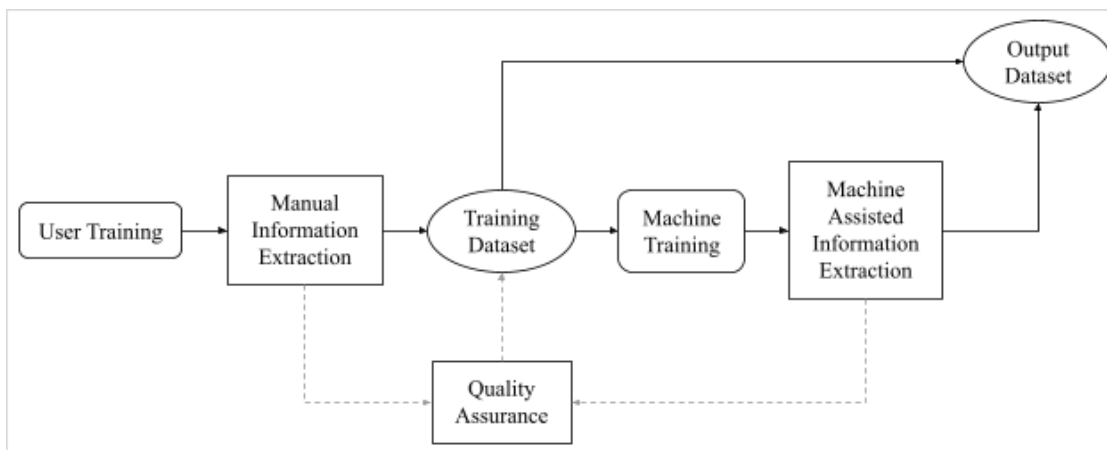


FIGURE B.1: SCOPE's User-oriented Learning-based System (repeat)

- *User Training*: Users of the SCOPE tool are trained to extract information based on the rules and needs of their project. For example, on the geoP-arsing Team at William & Mary, an undergraduate researcher might be

trained to identify and parse events of Chinese government-related infrastructure investment or development in Latin America and the Caribbean from open-source news articles.

- *Manual Information Extraction:* Based on their training, users then begin working manually to extract and parse information using the SCOPE tool.
- *Training Dataset:* The output of the Manual Information Extraction then forms the training dataset from which the model will learn. Some projects may retain a number of researchers working manually to collect data even after a machine learning model achieves high accuracy levels, so this dataset continually grows to fine tune the model. As with other methods of machine learning, the model will not be in danger of becoming hypertuned unless the researchers provide it with unrepresentative input data (Culotta et al., 2006). And since most projects will be domain-specific, a hypertuned model may not present limitations.
- *Machine Training:* The model(s) is continuously trained based on the data collected by and quality assured by the manual users.
- *Machine Assisted Information Extraction:* Based on the training dataset and any other parameters, the model(s) automatically extracts and parses information in a similar fashion to the manual users.
- *Quality Assurance:* A subsection of the data (between 0-100%) is subjected to one or more rounds of quality assurance. Here, any inaccuracies performed by the previous researcher(s) is corrected. If the data originated from the machine assisted side of the framework, then the model is corrected. The quality assured data then joins the training dataset.
- *Output Dataset:* All of the data processed in the SCOPE tool can then be exported for use.

## B.2 Modules, Classes, Suites, and Frameworks

The SCOPE tool is developed in Python’s Django web framework, which uses a system of models, views, and templates to define how data is stored, processed, and interacted with in various parts of the tool. Django proved to be especially useful for developing the tool as it allows for the easy organization and development of functionalities independently of each other. It also wraps database setup and management, data retrieval and editing, and other tasks into the same programmatic environment. All of the code is accessible at <https://github.com/wmgeolab/scope>.

To maximize flexibility while benefiting from user-oriented learning, the SCOPE tool is developed in independent, optional “modules” which are linked together by their expected inputs and outputs. Each module is a grouping of functionalities which enable users to process or interact with the data in some specific, predetermined manner. All modules also belong to one of four “classes”: Administration, Pipelines, Workflows, or Analytics. Modules which serve similar purposes are grouped together as “suites.” The collection of modules that someone chooses to include in their SCOPE tool is called a “framework” (i.e., the modules necessary to operate a specific project). One can think of the SCOPE tool as a massive factory in which administration modules are the contracts which establish the factory, pipelines are the trucks which bring in building materials and resources, workflows are the machines which transform those materials into products, and analytics are the finishing touches which add further customizations to those products. Figure B.2 illustrates the modular structure of the SCOPE tool. Existing modules are denoted with an asterisk (\*). Images of existing modules and concept art of modules to be developed in the next iteration of the SCOPE tool can be found in Appendix C.

The flexibility offered through SCOPE’s modularized approach enables users to tailor their framework to only include the modules which their project requires (Ahn, 2006). Whereas geoParsing requires an IE system which will transform unstructured text into structured events data (both Extracting and Parsing), another project may only need to identify relevant information and store it as unstructured events data (only Extracting). Some projects, such as our own



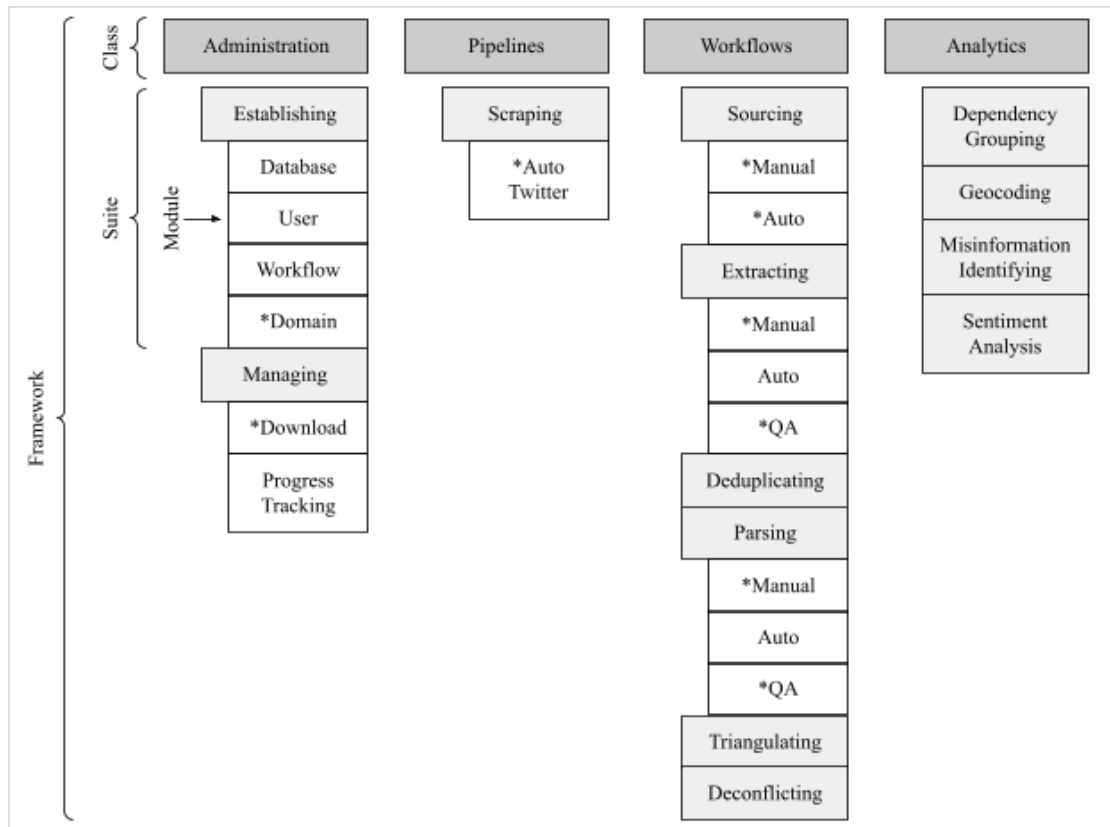


FIGURE B.2: Diagram of SCOPE's Modules in a Comprehensive Framework

coding efforts at William & Mary, may also benefit from insights into the intermediary stages of the IE process, as understanding which chunks of source text evidence which structured events can be incredibly valuable to the mission of identifying Chinese and Russian foreign activities and separating them from misinformation.

Aside from the Administration modules which must be included, a framework is highly flexible and can include different combinations of Pipelines, Workflows, and Analytics modules. The only requirement for which modules are included in a framework is that the inputs and outputs of modules are compatible. For Pipelines modules, data needs to match the format in the database which will then be queried for processing in Workflows modules. In Workflows modules, the Extracting Modules expect a source document from the Sourcing Modules, and the Parsing Modules expect an unstructured text extract from the Extracting Modules. In Analytics modules, expected inputs will also differ based

on whether it is expecting a source document, a shorter extract, or structured events data.

Administration modules are grouped into two suites, “Establishing modules” (Database, User, Workflow, and Domain), which set up the core functionalities of the SCOPE tool, and “Managing modules” (Download and Progress Tracking), which aid users in tracking the progress of their tasks. While these modules do not handle IE functionalities, each tackles an otherwise difficult obstacle for a user who might not be familiar with designing a database or creating user authorities from scratch. A tenet of user-oriented learning-based IIEEs is that they decrease the computational and data engineering expertise needed to create novel IE systems (Cardie and Pierce, 1998). These models can be explicitly defined as:

- *Database Module*: The first step is to establish a place where and define how all project data will be stored. In this module, a project manager is able to define the necessary databases and models for their SCOPE workflow. They can connect to an existing empty database or create a new database, either locally or hosted elsewhere to suit their project’s storage needs. They simply pass a local file path or URL/credentials. The project manager can either use the default SCOPE workflow models if they are creating an events dataset or design models tailored to their project’s needs. They can also store connection credentials for any other databases which their workflow will need to access, such as a database of sources which is grown outside of their SCOPE framework.
- *User Module*: After establishing the database, users need to be added to the project. In this module, a project manager is able to set up the responsibilities (modules each user has access to) and authorities (tasks each user can perform in a given module). Users are added by their GitHub usernames.
- *Workflow Module*: In this module, a project manager is able to set up the modules included in their SCOPE workflow, and the specifications of each. All administration modules are required. Pipelines, Workflows, and Analytics modules can be added based on the project’s needs. In Workflows

and Analytics modules (which are further distributed across Auto, Manual, and Quality Assurance) the percentage of data that is processed manually or with machine assistance can be set, as well as the percentage of data which is quality assured.

- *\*Domain Module:* In this module, a project manager is able to set up the domain-specific fields and codes for their information extraction schema. EE systems built around event trigger detection could incorporate a user-oriented system for developing event triggers and types (Tong et al., 2020).
- *\*Download Module:* Here, a project manager is able to export data from the SCOPE tool.
- *\*Progress Tracking Module:* Here, users with access can track the progress of the project, including breakdowns of work efficiency and statistics describing the data collected so far.

Another class of modules - Pipeline modules - are those which bring data into the database. They currently all belong to one suite, Scraping modules, which enable users to call REST APIs to query and import data from various sources. These modules are also supplemented by auxiliary programs we are developing outside of the SCOPE tool to constantly pull sources of information from Twitter and GDELT. These modules currently include:

- *\*Auto Twitter Scraping Module:* Through this module, any user with appropriate access is able to run a function which scrapes Twitter using its REST API to pull Tweets which match search criteria into an existing database model for sources. There is substantial literature which indicates the value of scraping online social networks for event extraction (Goswami and Kumar, 2016; Ritter, Etzioni, and Clark, 2012).

Workflow modules are those which make up the core of the information extraction functionalities of the SCOPE tool, whereas the other modules perform tasks which are in support of the information extraction task. Workflow modules transform one form of text-based data extracted from some online source into another form of text-based data. Most commonly, this will transform unstructured text from online news articles into structured observations in a dataset.

Workflows modules belong to several suites: Sourcing, Extracting, Deduplicating, Parsing, Triangulating, and Deconflicting. Each of these suites possesses Auto, Manual, and Quality Assurance modules to benefit from the comparative advantages of human and automated processes. In this first iteration of the SCOPE tool, only the Sourcing, Extracting, and Parsing classes are developed for basic within-source event extraction:

- *\*Sourcing (Manual)*: Through this module, any user with appropriate access is able to manually import sources into the respective sourcing model of their workflow database. Sources can be added individually via URL or by importing a CSV file with source information.
- *\*Sourcing (Auto)*: Here, any user with appropriate access is able to run a function which queries their database of sources and pulls the results into the respective sourcing model of their workflow database.<sup>1</sup> This functionality was developed independently of the Sourcing (Manual) Module as it may not be the case that every project has or needs access to an auxiliary database of sources.
- *\*Extracting (Manual)*: Users are able to manually process a percentage of sources (currently 100%) for event detection and relevant text for each event is sent as extracts to the next module. What constitutes a relevant event is based on the schema of the project and users' training. For example, a relevant event for geoParsing might be the signing of an agreement to build a new Chinese hydroelectric dam in Ecuador in 2018. All non-redundant descriptive text related to this event in the online article is grouped together, unaltered, to create an extract. The inclusion of Extracting modules recognizes the findings of abundant literature at the across-sentence level and document level of event extraction (Duan, He, and Zhao, 2017; Naughton, Kushmerick, and Carthy, 2006; Thompson et al., 2017; Du and Cardie, 2020; Du, Rush, and Cardie, 2020; Huang and Peng, 2020). The Extracting module enables users to bundle together all of the 5W1H information and meta-knowledge of an event into a single chunk of text. Multiple extracts allow for the extraction of multiple events from the

---

<sup>1</sup>For now, this module is not as "automatic" as the naming convention would suggest, as it still requires interaction from the user.

same document, or even the same sentence (Liu, Luo, and Huang, 2018; Li et al., 2019). An “Auto Assist” button enables users to run the same machine learning model used in the Auto module here and immediately make edits to the output. The current “Auto Assist” button calls a basic function which simply returns the few sentences before and after each occurrence of a predefined trigger word. If a source has no relevant information, then it will have nothing sent forward - creating a Sourcing (QA) Module would have been redundant for this reason.

- *Extracting (Auto)*: This module would serve as the housing infrastructure for a machine learning model to replicate the process of the Extracting (Manual) Module without direct interaction with the user. Here, a project manager would select which model(s) to use to perform the extraction task.
- *\*Extracting (QA)*: Users with appropriate access are able to quality assure a percentage (currently 100%) of the extracts produced in the Manual and Auto modules. Once an extract has been quality assured by a number of independent users (currently 1), it can either be sent forward to the next module if accurate, or sent back to the Manual module. Future iterations of the SCOPE tool might design quality assurance differently to preserve the machine learning principles of user-oriented learning while preserving the user friendliness of the tool.
- *\*Parsing (Manual)*: Users are able to manually parse a percentage of extracts (currently 100%) into structured events data with fields established in the Database and Domain Modules. This is where the 5W1H (who, what, when, where, why, and how) topics of the event are addressed and the data is transformed into a format which is more accessible to most computer software once exported from the SCOPE tool (e.g., a CSV file to be examined in Microsoft Excel). Recognizing that some extracts may contain details on several events, multiple can be parsed from a single event (Liu, Luo, and Huang, 2018). For example, a sentence may contain a list which references a dam, a highway, and a school all being built by China in Ecuador. This would not be easily separable in the current Extracting modules. An “Auto Assist” button would enable users to run the same

machine learning model used in the Auto module here and immediately make edits to the output.

- *Parsing (Auto)*: This module would serve as the housing infrastructure for a machine learning model to replicate the process of the Parsing (Manual) Module without direct interaction with the user. Here, a project manager would select which model(s) to use to perform the parsing task.
- *\*Parsing (QA)*: Users with appropriate access are able to quality assure a percentage (currently 100%) of the structured events produced in the Manual and Auto modules. Once an extract has been quality assured by a number of independent users (currently 1), it can either be sent forward to the next module if accurate, or sent back to the Manual module. Future iterations of the SCOPE tool might design quality assurance differently to preserve the machine learning principles of user-oriented learning while preserving the user friendliness of the tool.

There are additional Workflow modules which were considered for development but not included in the first iteration of the SCOPE tool. A Deduplicating Module was considered for resolving situations when information on one event is incorrectly grouped into two separate extracts instead of a single one. In most cases, this functionality seems unnecessary if the Extracting modules are efficacious, though, may be valuable for extracting information from documents which are very long. There is substantial literature which indicates that document-level EE outperforms sentence-level EE given the availability of more contextual details scattered beyond a single sentence, but document length sometimes constrains computational feasibility (Du and Cardie, 2020; Duan, He, and Zhao, 2017; Huang and Peng, 2020; Thompson et al., 2017; Yang and Mitchell, 2016). A Triangulating Module was considered for linking related information which has been extracted from several sources, or an across-document level of analysis. Most IE literature focuses on either sentence-level or document-level extraction processes, and this would be the first user-oriented learning IIEE to systematically approach across-document-level extraction to the best of the author's knowledge (Naughton, Kushmerick, and Carthy, 2006; Du, Rush, and Cardie, 2020; Wan and Yang, 2008). Such a module would be especially useful for IE processes which may be susceptible to misinformation, such

as geoParsing's work tracking Chinese and Russian foreign activities. A De-conflicting Module was considered for resolving conflicting data describing the same event at the across-document level (Naughton, Kushmerick, and Carthy, 2006; Agerri et al., 2016).

Analytics modules are those which perform additional tasks for the data deliverables which are included in the information extraction workflow. Modules which were considered for development include Dependency Grouping Modules, which would group and rank events based on how they relate to each other, and Geocoding Modules, which would assign coordinate data to the events. Dependency Grouping would address nested event structures where, for example a "crime" event can cause an "investigation" event and eventually an "arrest" event (McClosky, Surdeanu, and Manning, 2011; Li et al., 2019). Another Analytics module which was considered was a Misinformation Identifying Module, which would assess the information in source documents for potential misinformation. Similarly, a Sentiment Analysis Module would be developed to assess the type of words used to describe events in various online news sources.

## Appendix C

# Graphics of the Scope Tool

SCOPE Tool -- Complete List of Modules

### Administration

- Establishing
- Database +
- Users +
- Workflow +
- Domain \*
- Managing
- Download \*
- Progress Tracking

### Pipelines

- Scraping
- Auto Twitter Scraping \*

### Workflows

- Sourcing
- Manual \*
- Auto \*
- Extracting
- Manual \*
- Auto
- Quality Assurance \*
- Deduplicating
- Parsing
- Manual \*
- Auto
- Quality Assurance \*
- Triangulating
- Deconflicting

### Analytics

- Dependency Grouping
- Geocoding
- Misinformation Identifying
- Sentiment Analysis

\* = created

+ = concept art designed



Database Module

\*concept art for next iteration of SCOPE

### Database Setup

Connect to an existing database.

Successfully connected to database

[Do not already have a database?](#)

Set up models.

### Local Database Connection

Browse your local computer for a database.

No path chosen

### Remote Database Connection

Enter the URL and login credentials for your database.

Host:

Database:

User:

Password:





User Module

\*concept art for next iteration of SCOPE

**Add Collaborator**

Enter GitHub username or email address

Only team members can be added as collaborators.

Collaborator	Extracting	Parsing	QA Parsing		
 wgpettus	Viewer	Contributor	Contributor	<input type="button" value="change"/>	<input type="button" value="delete"/>
 mmfritz	Contributor	Contributor	Viewer	<input type="button" value="change"/>	<input type="button" value="delete"/>
 micrittenden	Admin	Contributor	Contributor	<input type="button" value="change"/>	<input type="button" value="delete"/>
 dsmlerrunfol	Admin	Admin	Admin	<input type="button" value="change"/>	<input type="button" value="delete"/>

# micrittenden

---

**Manual Extracting:**

**Manual Parsing:**

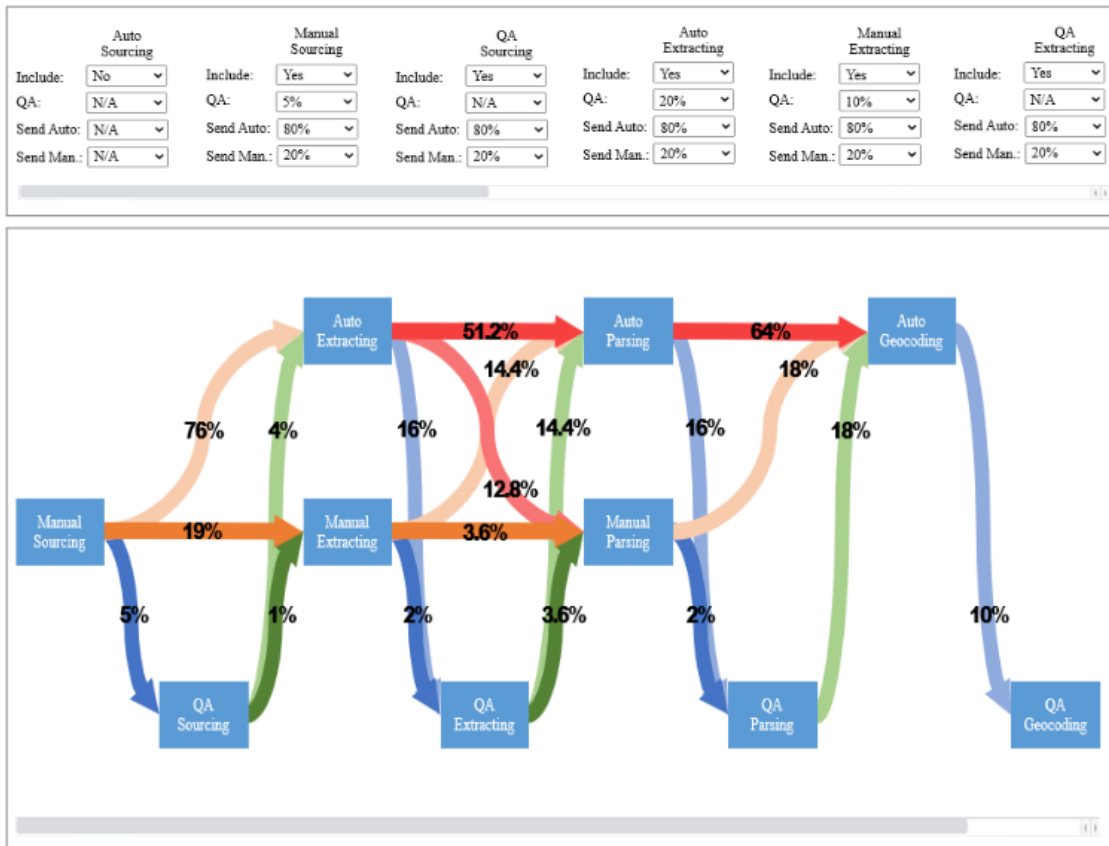
**QA Parsing:**

**Auto Geocoding:**

**QA Geocoding:**

## Workflow Module

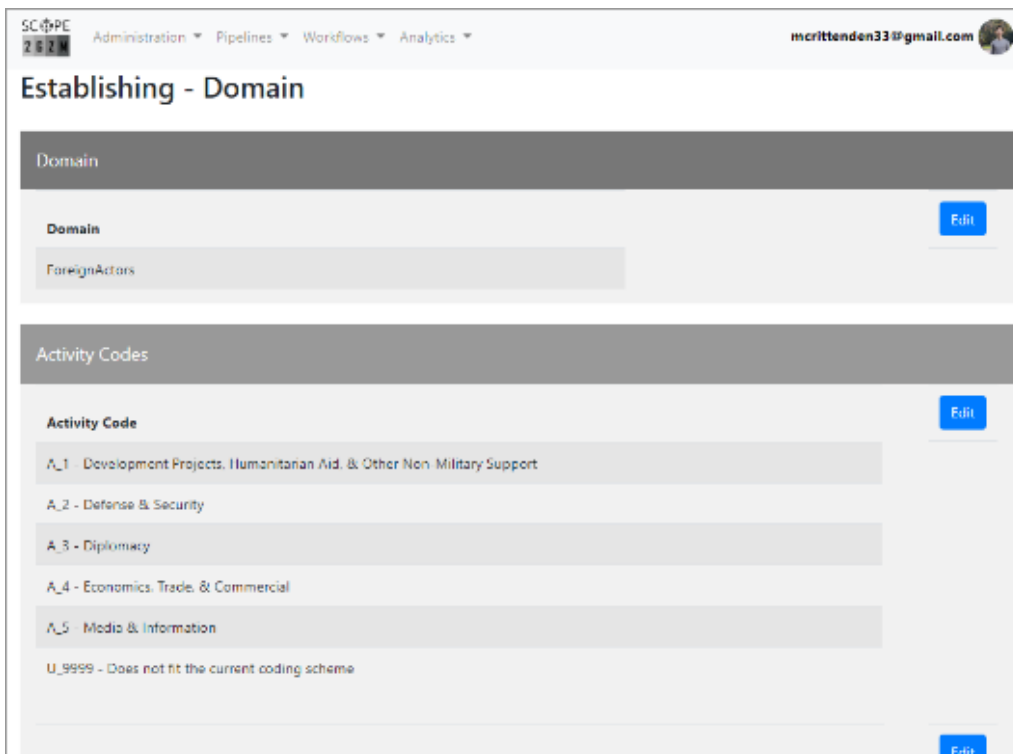
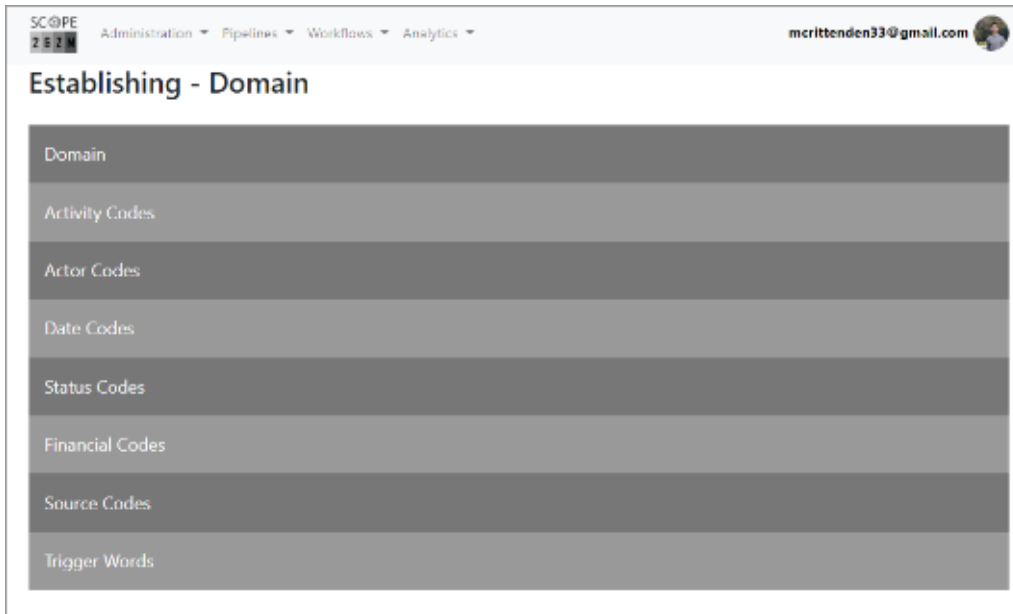
\*concept art for next iteration of SCOPE



Specifications for the above concept art:

- Send 5% of Manual Sources to QA Sourcing
- Send 20% of Auto Extracting to QA Extracting
- Send 10% of Manual Extracting to QA Extracting
- Send 20% of Auto Parsing to QA Parsing
- Send 10% of Manual Parsing to QA Parsing
- Send 10% of Auto Geocoding to QA Geocoding
- Send 80% of Manual Sourcing to Auto Extracting
- Send 20% of Manual Sourcing to Manual Extracting
- Send 80% of Auto Extracting to Auto Parsing
- Send 20% of Auto Extracting to Manual Parsing
- Send 80% of Manual Extracting to Auto Parsing
- Send 20% of Manual Extracting to Manual Parsing
- Send 100% of Auto Parsing to Auto Geocoding
- Send 100% of Manual Parsing to Auto Geocoding

### Domain Module



Download Module

## Download

To download data from a module, select one of the options below:

Sources

Extracts

Activities

Auto Twitter Scraping Module

\*concept art for next iteration of SCOPE

### Auto Twitter Scraping

To scrape from Twitter, input query specifications in the inputs below:

Primary Keywords:

Secondary Keywords:

Tertiary Keywords:

Start Date:

End Date:

Submit

January 2019						
Su	Mo	Tu	We	Th	Fr	Sa
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

## Manual Sourcing Module

Sourcing				
Source ID	Source Code	URL	Added	Action
1	S_4 - Media reports	<a href="https://www.bbc.com/news/world-asia-55889565">https://www.bbc.com/news/world-asia-55889565</a>	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
2	S_4 - Media reports	<a href="https://www.bbc.com/news/world-asia-55913947">https://www.bbc.com/news/world-asia-55913947</a>	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
3	S_4 - Media reports	<a href="https://www.bbc.com/news/world-asia-55851052">https://www.bbc.com/news/world-asia-55851052</a>	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
4	S_4 - Media reports	<a href="https://www.bbc.com/news/world-africa-55934277">https://www.bbc.com/news/world-africa-55934277</a>	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
5	S_4 - Media reports	<a href="https://www.bbc.com/news/technology-56103921">https://www.bbc.com/news/technology-56103921</a>	Feb. 17, 2021, 10:19 p.m.	<a href="#">View</a>

HTML Source	
Source code:	S_4 - Media reports
Source url:	<a href="https://www.bbc.com/news/world-africa-55934277">https://www.bbc.com/news/world-africa-55934277</a>
Source html:	<pre>&lt;!DOCTYPE html&gt; &lt;html lang="en-GB" class="no-js"&gt; &lt;head&gt; &lt;meta charset="utf-8" /&gt; &lt;meta name="viewport" content="width=device-width, initial-scale=1" /&gt; &lt;title data-rh="true"&gt; Nigerian separatist Nnamdi Kanu&amp;#x27;s Facebook account removed for hate speech - BBC News&lt;/title&gt; &lt;meta data-rh="true" name="description" content="Nnamdi Kanu posted video of a militia attack amid a bitter dispute between cattle herders and farmers." /&gt; &lt;meta data-rh="true" name="theme-color" content="#FFFFFF" /&gt; &lt;meta data-rh="true" property="article:author" content="https://www.facebook.com/bbcnews/" /&gt; &lt;meta data-rh="true" property="article:section" content="Africa" /&gt; &lt;meta data-rh="true" property="fb:admins" content="100004154058350" /&gt; &lt;meta data-rh="true" property="fb:app_id" content="1609039196070050" /&gt; &lt;meta data-rh="true" property="fb:pages" content="1143803202301544,317278538359186,1392506827668140,742734325867560,185246968166196,156060587793370, 137920769558355,193435954068976,21263239760,156400551056385,929399697073756,154344434967,228735667216,8075 Nigerian separatist Nnamdi Kanu's Facebook account removed for hate speech Published image copyrightGetty Images image captionNnamdi Kanu wants independence for south-eastern Nigeria Facebook says it removed the page of Nigerian separatist leader Nnamdi Kanu for violating its rules on harm and hate speech. Mr Kanu's page was removed for repeated violation of its community rules, the social networking site told the BBC. He had posted a video of a militia group attacking and killing cattle in a herders' settlement. He also used the live broadcast to accuse herders of destroying farmlands in eastern Nigeria. The conflict between herders and other groups is currently one of Nigeria's hottest political issues. Mr Kanu leads the Indigenous People of Biafra (Ipub), which campaigns for independence for Nigeria's south-eastern region, where the ethnic Igbo people form the majority. The herders are mostly from the northern Fulani community. Mr Kanu, who also has British nationality, used his Facebook page as a key platform to communicate with his followers around the world. The account was blocked on Tuesday. 'Suppressing the truth' The militia carrying out the attack in the video he posted are</pre>
Source text:	Nigerian separatist Nnamdi Kanu's Facebook account removed for hate speech Published image copyrightGetty Images image captionNnamdi Kanu wants independence for south-eastern Nigeria Facebook says it removed the page of Nigerian separatist leader Nnamdi Kanu for violating its rules on harm and hate speech. Mr Kanu's page was removed for repeated violation of its community rules, the social networking site told the BBC. He had posted a video of a militia group attacking and killing cattle in a herders' settlement. He also used the live broadcast to accuse herders of destroying farmlands in eastern Nigeria. The conflict between herders and other groups is currently one of Nigeria's hottest political issues. Mr Kanu leads the Indigenous People of Biafra (Ipub), which campaigns for independence for Nigeria's south-eastern region, where the ethnic Igbo people form the majority. The herders are mostly from the northern Fulani community. Mr Kanu, who also has British nationality, used his Facebook page as a key platform to communicate with his followers around the world. The account was blocked on Tuesday. 'Suppressing the truth' The militia carrying out the attack in the video he posted are
Source date:	2021-02-04 00:00:00
Date added:	2021-02-05 14:31:15
Current user:	-----
Current status:	unprocessed

### Add New HTML Source

Source code:

Source url:

Source date:

### Import New HTML Sources

Browse your local computer for a data file (.csv) to upload.

No file chosen



Auto Sourcing Module

\*concept art for next iteration of SCOPE

### Auto Sourcing

[Auto Import](#)

Source ID	Source Code	URL	Added	Action
1	S_4 - Media reports	https://www.bbc.com/news/world-asia-55889565	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
2	S_4 - Media reports	https://www.bbc.com/news/world-asia-55913947	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
3	S_4 - Media reports	https://www.bbc.com/news/world-asia-55851052	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
4	S_4 - Media reports	https://www.bbc.com/news/world-africa-55934277	Feb. 5, 2021, 2:31 p.m.	<a href="#">View</a>
5	S_4 - Media reports	https://www.bbc.com/news/technology-56103921	Feb. 17, 2021, 10:19 p.m.	<a href="#">View</a>

### Auto Sourcing

To import sources from the master database, fill out the fields below and click Run.

Primary Keywords:

Secondary Keywords:

Tertiary Keywords:

Start Date:

End Date:

< January 2019 >

Su	Mo	Tu	We	Th	Fr	Sa
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

[Run](#)

### Manual Extracting Module

**Extraction**

To extract information from a source, select a source from the list below:

Source ID	Source Code	URL	Extracts	Action
4	S_4 - Media reports	https://www.bbc.com/news/world-africa-55934277	0	<button>Checkout</button>
5	S_4 - Media reports	https://www.bbc.com/news/technology-56103921	0	<button>Checkout</button>

SCOPE 267M Administration Pipelines Workflows Analytics mccrittenden33@gmail.com

## Checking out source for extraction Release Save Finish

**Source:**

Source ID: 150  
 Source Code: SCOPE\_S\_1 - Online news article  
 URL: <https://www.bbc.com/news/world-asia-china-56607815>

---

**Extracts:** Auto Assist

**Text:**  **Delete:**

## Quality Assurance Extracting Module

SC@PE 262M Administration Pipelines Workflows Analytics mcrittenden33@gmail.com

### Extracting - QA

To quality assure an extract, select from the list below.

[Or click here to view a complete list of extracts and their progress.](#)

Extract ID	Source ID	Text	Action
517	105	Past the rival billboards, Russian-sponsored radio stations, school classes, cartoons, film screenings and beauty contests are all jostling for locals' attention. French and Russian propaganda posters jostle for attention in downtown Bangui.	QA
519	105	One of CAR's top rebel commanders, MPC's Mahamat Alkatim, who in August 2019 resigned as special military adviser - a government position granted as a concession to armed groups under this year's peace treaty, which was partly overseen by Russia	QA
520	105	he station's EU-backed rival is "Radio Ndeke Luka", a popular outlet that describes itself as "the radio of all Central Africans", with reports covering anything from government corruption to women's rights and negotiations with rebels. Beyond broadcasting, Russia puts on "Friendship Lessons" in CAR's schools, involving specially curated history courses and boxes of free "WE LOVE RUSSIA" t-shirts. Moscow's money has paid for a new youth football tournament launched on Saturday as well as gym equipment draped in Russian flags. Pro-Russian banners at these events are written not in French—an official language from the colonial era—but in the vernacular, Sango.	QA
572	186	Russia has delivered light arms to CAR's security forces this year and deployed hundreds of military and civilian instructors to train them.	QA
577	185	ADDIS ABABA (Thomson Reuters Foundation) - In a damp office at Ethiopia's Addis Ababa University, doctoral student Hailu Geremew fantasizes about working on the nuclear reactor his country is now	QA


SC@PE 262M Administration Pipelines Workflows Analytics mcrittenden33@gmail.com

### Extracting - QA

#### Checking out extract for parsing

Release Send Back Send Forward

**Extract:**



Extract ID: 517  
 Source ID: 105  
 Source URL: <https://www.telegraph.co.uk/news/2019/09/14/inside-russias-soft-power-battle-west-civil-war-ravaged-central/>  
 Extract Text: Past the rival billboards, Russian-sponsored radio stations, school classes, cartoons, film screenings and beauty contests are all jostling for locals' attention. French and Russian propaganda posters jostle for attention in downtown Bangui.



## Quality Assurance Parsing Module

SC@PE 2.8.2 Administration Pipelines Workflows Analytics mrcrittenden33@gmail.com

## Parsing - QA

To quality assure an activity, select from the list below:  
[Or click here to view a complete list of activities and their progress.](#)

Activity ID	Extract ID	Source ID	Extract Text	Action
2	15	6	Mexico on Monday auctioned eight out of ten deep water oil and gas blocks up for grabs in the Gulf of Mexico, and scored a joint venture for a major crude field in the most hotly-anticipated round of the country's energy opening so far. China's Offshore Oil Corporation took two of the eight blocks.	QA
3	9	18	Bolivian Energy Minister Rafael Alarcon said Friday that Chinese company Sinohydro has won a tender to build the Ivirizu hydroelectric dam.	QA
4	16	6	Almost 1.2 billion of those were areas rich in most-valuable light and super light crude secured by China Offshore, a unit of the Chinese giant CNOOC, in the Perdido Fold Belt, where output on the U.S. side of the formation has been booming for years.	QA
5	17	10	The government of the People's Republic of China, in conjunction with the Export-Import Bank of China, has entered a financial arrangement with T&T to provide \$210 million for the completion of infrastructural (remedial) works on the National Academy for the Performing Arts (NAPA) in Port-of-Spain.	QA
6	10	18	The Ivirizu project will be developed over four years, with an investment of over 550 million U.S. dollars to bring 279.9 megawatts to the national grid once completed. The complex will have two hydroelectric power plants -- Sehuenas, which will produce 188.62MW, and Pelton, which will generate 64MW.	QA
8	23	19	China's West Pacific Petrochemical Corp, or WEPEC, is set to export 900,000 barrels of gasoline to Mexico this month amid a swelling glut of the motor fuel at home, according to an industry source with direct knowledge of the matter.	QA


SC@PE 2.8.2 Administration Pipelines Workflows Analytics mrcrittenden33@gmail.com

## Parsing - QA

### Checking out activity for parsing

Release Send Back Send Forward

Activity:



Activity ID: 2  
 Extract ID: 15  
 Source ID: 6  
 Source URL: <https://www.reuters.com/article/us-mexico-oil-auction-first/china-stakes-claim-in-mexico-oil-opening-at-deep-water-auction-idUSKBN13U212>  
 Extract Text: Mexico on Monday auctioned eight out of ten deep water oil and gas blocks up for grabs in the Gulf of Mexico, and scored a joint venture for a major crude field in the most hotly-anticipated round of the country's energy opening so far. China's Offshore Oil Corporation took two of the eight blocks.  
 Activity Subcode: A\_1004 - Development Projects, Humanita... - Energy & Technology (General)  
 Actor Code: C\_16 - Organization, Consortium, or Group  
 Actor Name: China's Offshore Oil Corporation  
 Actor Role: R\_10 - Implementer  
 Activity Date: Dec. 5, 2016  
 Date Code: D\_1 - Proposed - formal  
 Status Code: E\_1 - Independent event  
 Financial Code: None  
 Dollar Amount: None  
 Locations: None  
 Parser Notes: None

# Bibliography

- ACLED (2019). *ACLED Methodology*. URL: [https://acleddata.com/acleddatanew/wp-content/uploads/dlm\\_uploads/2019/04/Methodology-Overview\\_FINAL.pdf](https://acleddata.com/acleddatanew/wp-content/uploads/dlm_uploads/2019/04/Methodology-Overview_FINAL.pdf).
- (2020). *FAQs: ACLED Sourcing Methodology*. URL: [https://acleddata.com/acleddatanew/wp-content/uploads/dlm\\_uploads/2020/02/FAQs\\_ACLED-Sourcing-Methodology.pdf](https://acleddata.com/acleddatanew/wp-content/uploads/dlm_uploads/2020/02/FAQs_ACLED-Sourcing-Methodology.pdf).
- (2021). *The Armed Conflict Location & Event Data Project: Bringing clarity to crisis*. URL: <https://acleddata.com/#/dashboard>.
- Agerri, Rodrigo et al. (2016). “Multilingual Event Detection using the News-Reader Pipelines”. In: *Proceedings of the Cross-Platform Text Mining and Natural Language Processing Interoperability workshop*. Portoroz, Slovenia, pp. 42–46.
- Ahn, David (2006). “The stages of event extraction”. In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Sydney, Australia, pp. 1–8.
- Broman, Karl W. and Kara H. Woo (2018). “Data Organization in Spreadsheets”. In: *The American Statistician* 72.1, pp. 2–10. DOI: 10.1080/00031305.2017.1375989. eprint: <https://doi.org/10.1080/00031305.2017.1375989>. URL: <https://doi.org/10.1080/00031305.2017.1375989>.
- Campos, David et al. (2014). “Egas: a collaborative and interactive document curation platform”. In: *The Journal of Biological Databases and Curation*. ISSN: 1758-0463. DOI: 10.1093/database/bau048. eprint: <https://academic.oup.com/database/article-pdf/doi/10.1093/database/bau048/8245970/bau048.pdf>. URL: <https://doi.org/10.1093/database/bau048>.
- Cardie, Claire and D. Pierce (1998). *Proposal for an Interactive Environment for Information Extraction*. Tech. rep. Ithaca, NY: Cornell University, pp. 1–12. eprint: <http://www.cs.cornell.edu/home/pierce/papers/cucstr98.pdf>. URL: <http://www.cs.cornell.edu/home/pierce/papers/cucstr98.pdf>.

- Cejuela, Juan Miguel et al. (2014). "tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles". In: *The Journal of Biological Databases and Curation*, pp. 1–8. ISSN: 1758-0463. DOI: [10.1093/database/bau033](https://doi.org/10.1093/database/bau033). eprint: <https://academic.oup.com/database/article-pdf/doi/10.1093/database/bau033/8245396/bau033.pdf>. URL: <https://doi.org/10.1093/database/bau033>.
- Chahal, Manoj (2014). "Information Retrieval using Jaccard Similarity Coefficient". In: *International Journal of Computer Trends and Technology* 36.3, pp. 140–142. ISSN: 2231-2803. DOI: [10.14445/22312803/IJCTT-V36P124](https://doi.org/10.14445/22312803/IJCTT-V36P124). eprint: <https://www.ijcttjournal.org/2016/Volume36/number-3/IJCTT-V36P124.pdf>. URL: <http://www.ijcttjournal.org/archives/ijctt-v36p124>.
- Chan, Yee Seng et al. (2019). "Rapid Customization for Event Extraction". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy, pp. 31–36.
- Chen, Yubo et al. (2017). "Automatically Labeled Data Generation for Large Scale Event Extraction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, pp. 409–419.
- Culotta, Aron et al. (2006). "Corrective feedback and persistent learning for information extraction". In: *Artificial Intelligence* 170.14, pp. 1101–1122. ISSN: 0004-3702. DOI: <https://doi.org/10.1016/j.artint.2006.08.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0004370206000762>.
- Du, X., Alexander M. Rush, and Claire Cardie (2020). "Document-level Event-based Extraction Using Generative Template-filling Transformers". In: *ArXiv abs/2008.09249*, pp. 1–13.
- Du, Xinya and Claire Cardie (2020). "Document-Level Event Role Filler Extraction using Multi-Granularity Contextualized Encoding". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8010–8020.
- Duan, Shaoyang, Ruifang He, and Wenli Zhao (2017). "Exploiting document level information to improve event detection via recurrent neural networks". In: *Proceedings of the 8th International Joint Conference on Natural Language Processing*. Taipei, Taiwan, pp. 352–361.
- GDELT (2015). *GDELT 2.0: Our Global World in Realtime*. URL: <https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>.

- Gomaa, Wael H. and Aly A. Fahmy (2013). "A survey of text similarity approaches". In: *International Journal of Computer Applications* 68.13, pp. 13–18. ISSN: 0975 – 8887. DOI: [10.5120/11638-7118](https://doi.org/10.5120/11638-7118). eprint: <https://research.ijcaonline.org/volume68/number13/pxc3887118.pdf>. URL: <http://dx.doi.org/10.5120/11638-7118>.
- Goswami, Anuradha and Ajey Kumar (2016). "A survey of event detection techniques in online social networks". In: *Social Network Analysis and Mining* 6.107, pp. 1–25. ISSN: 1869-5469. DOI: [10.1007/s13278-016-0414-1](https://doi.org/10.1007/s13278-016-0414-1). URL: <https://doi.org/10.1007/s13278-016-0414-1>.
- Grishman, Ralph, Silja Huttunen, and Roman Yangarber (2002). "Real-time event extraction for infectious disease outbreaks". In: *Proceedings of 2nd International Conference on Human Language Technology Research*. San Francisco, CA, pp. 366–369.
- Halterman, Andrew et al. (2017). "Adaptive scalable pipelines for political event data generation". In: *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2879–2883. DOI: [10.1109/BigData.2017.8258256](https://doi.org/10.1109/BigData.2017.8258256).
- Hogenboom, Frederik et al. (2016). "A Survey of Event Extraction Methods from Text for Decision Support Systems".
- Huang, Kung-Hsiang and Nanyun Peng (2020). "Efficient End-to-end Learning of Cross-event Dependencies for Document-level Event Extraction". In: *ArXiv abs/2010.12787*, pp. 1–10.
- Huttunen, Silja et al. (2013). "Predicting Relevance of Event Extraction for the End User". In: *Multi-source, Multilingual Information Extraction and Summarization*. Ed. by Thierry Poibeau et al. Heidelberg, Germany: Springer, pp. 163–176.
- Li, Wei et al. (2019). "Joint Event Extraction Based on Hierarchical Event Schemas From FrameNet". In: *IEEE Access* 7, pp. 25001–25015. DOI: [10.1109/ACCESS.2019.2900124](https://doi.org/10.1109/ACCESS.2019.2900124).
- Liu, Xiao, Zhunchen Luo, and Heyan Huang (2018). "Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 1247–1256.
- McClosky, David, Mihai Surdeanu, and Christopher Manning (2011). "Event Extraction as Dependency Parsing". In: *Proceedings of the 49th Annual Meeting*



- of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR.
- National Statistics, UK Office for (2020). *Global Database of Events, Language and Tone (GDELT) data quality note*. URL: <https://backup.ons.gov.uk/wp-content/uploads/sites/3/2020/01/Global-Database-of-Events-Language-and-Tone-GDELT-data-quality-note.pdf>.
- Naughton, Martina, Nicholas Kushmerick, and Joseph Carthy (2006). "Event Extraction from Heterogeneous News Sources". In: *Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis*, pp. 1626–1635.
- Pafilis, Evangelos et al. (2016). "EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation". In: *The Journal of Biological Databases and Curation*, pp. 1–7. ISSN: 0027-8424. DOI: [10.1093/database/baw005](https://doi.org/10.1093/database/baw005). eprint: <https://academic.oup.com/database/article-pdf/doi/10.1093/database/baw005/8222692/baw005.pdf>. URL: <https://doi.org/10.1093/database/baw005>.
- Pierce, D. and Claire Cardie (2001). *User-Oriented Machine Learning Strategies for Information Extraction: Putting the Human Back in the Loop*. Tech. rep. Ithaca, NY: Cornell University, pp. 1–2. URL: [https://www.researchgate.net/publication/2368251\\_User-Oriented\\_Machine\\_Learning\\_Strategies\\_for\\_Information\\_Extraction\\_Putting\\_the\\_Human\\_Back\\_in\\_the\\_Loop](https://www.researchgate.net/publication/2368251_User-Oriented_Machine_Learning_Strategies_for_Information_Extraction_Putting_the_Human_Back_in_the_Loop).
- Riedel, Sebastian and Andrew McCallum (2011). "Fast and Robust Joint Methods for Biomedical Event Extraction". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, pp. 1–12.
- Ritter, Alan, Oren Etzioni, and Sam Clark (2012). "Open Domain Event Extraction from Twitter". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Beijing, China, pp. 1104–1112. DOI: [10.1145/2339530.2339704](https://doi.org/10.1145/2339530.2339704). URL: <https://doi.org/10.1145/2339530.2339704>.
- Salam, Sayeed et al. (2018). "Distributed Framework for Political Event Coding in Real-Time". In: *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*, pp. 266–273. DOI: [10.1109/EECS.2018.00057](https://doi.org/10.1109/EECS.2018.00057).
- Sarawagi, Sunita (2007). "Information Extraction". In: *Foundations and Trends in Databases* 1.3, pp. 261–377. DOI: [10.1561/1500000003](https://doi.org/10.1561/1500000003).

- Taylor, David McD. et al. (2020). "Research skills and the data spreadsheet: A research primer for low- and middle-income countries". In: *African Journal of Emergency Medicine* 10, S140–S144. DOI: <https://doi.org/10.1016/j.afjem.2020.05.003>. URL: <https://www.sciencedirect.com/science/article/pii/S2211419X20300380>.
- Thompson, Paul et al. (2017). "Enriching news events with meta-knowledge information". In: *Language Resources and Evaluation* 51.2, pp. 409–438. ISSN: 1574-0218. DOI: [10.1007/s10579-016-9344-9](https://doi.org/10.1007/s10579-016-9344-9). URL: <https://doi.org/10.1007/s10579-016-9344-9>.
- Tong, Meihan et al. (2020). "Improving Event Detection via Open-domain Trigger Knowledge". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5887–5897.
- Venugopal, Deepak et al. (2014). "Relieving the Computational Bottleneck: Joint Inference for Event Extraction with High-Dimensional Features". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, pp. 831–843.
- Wan, Xiaojun and Jianwu Yang (2008). "Multi-document summarization using cluster-based link analysis". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Singapore, pp. 299–306. URL: <https://doi.org/10.1145/1390334.1390386>.
- Wang, Wei (2017). "Event Detection and Encoding from News Articles". PhD thesis. Blacksburg, VA: Virginia Polytechnic Institute and State University.
- Xiang, Wei and Bang Wang (2019). "A Survey of Event Extraction from Text". In: *IEEE Access* 7, pp. 1–29. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2019.2956831](https://doi.org/10.1109/ACCESS.2019.2956831). eprint: [https://www.researchgate.net/publication/337638438\\_A\\_Survey\\_of\\_Event\\_Extraction\\_From\\_Text](https://www.researchgate.net/publication/337638438_A_Survey_of_Event_Extraction_From_Text). URL: [https://www.researchgate.net/publication/337638438\\_A\\_Survey\\_of\\_Event\\_Extraction\\_From\\_Text](https://www.researchgate.net/publication/337638438_A_Survey_of_Event_Extraction_From_Text).
- Yang, Bishan and Tom Mitchell (2016). "Joint Extraction of Events and Entities within a Document Context". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, pp. 289–299.