Kingfisher

SYSTEMS, INC.

WHITE PAPER

VARYSS

Validated data, Analyzed by data scientists and analysts, using Research only made possible by a computational approach Yielding visual information including tables, graphs or charts, Submitted for further refinement before Sending to end user.

Prepared By:

Kingfisher Systems, Inc.

3110 Fairview Park Drive, Suite 1250

Falls Church, VA 22042

Phone: (703) 635-2951

Fax: (703) 820-7976

Email: varyss@kingfishersys.com

About Kingfisher's Computational Analytics Division

Kingfisher is a Capability Maturity Model Integration (CMMI) CMMI-DEV Level 3 appraised and International Organization for Standardization (ISO) 9001:2015 certified company. We integrate our CMMI practices into our Agile development methodology to help produce quality software output on a schedule appropriate for that type of development. We have over a decade's worth of experience solving complex problems using our Open Source Intelligence (OSINT) collection and analytics product, VARYSS, for various U.S. Government customers.

VARYSS has been operational for the last decade, collecting OSINT media in over 90 languages. VARYSS currently collects over 500,000 news articles a day from sources in every country with Internet access. Our modular architecture enables us to rapidly train NLP primitives and obtain higher-level statistical capabilities in languages that we have not previously exploited, increasing the pool of languages for which we have a deep understanding and sophisticated social and technical statistics. We are constantly expanding the features and capabilities of our VARYSS product to remain the market leader in OSINT collection and processing.

About VARYSS

Kingfisher's VARYSS product is the result of over a decade of experience using Artificial Intelligence (AI), Machine Learning (ML), complex statistical analysis, and modeling of large datasets to find unique and surprising answers to various problems for the Department of Defense (DoD), Department of Homeland Security (DHS), and the Intelligence Community (IC). We focus on ingesting unstructured OSINT and exploiting it with custom algorithms and deep neural methods to detect, identify, and classify structure within the data to create political time series. At Kingfisher there has been a focus on deep learning for time series, which enables us to produce a sophisticated real-time description of the continually evolving international information landscape. We have a proprietary list of over 65,000 news sites and 3,000 periodic journals and are actively expanding our collections to surpass the million-articles-a-day benchmark in 2019.

Product Work Flow

VARYSS is built on a modular infrastructure, using Linux servers, Mongo and PostgreSQL databases, and a mix of open-source, Commercial Off-The-Shelf (COTS), and proprietary tools for processing the data. We manage the Intellectual Property (IP) rights of our software aggressively, so that none of the integrated technologies limit our ability to use, sell, or distribute our VARYSS tool.

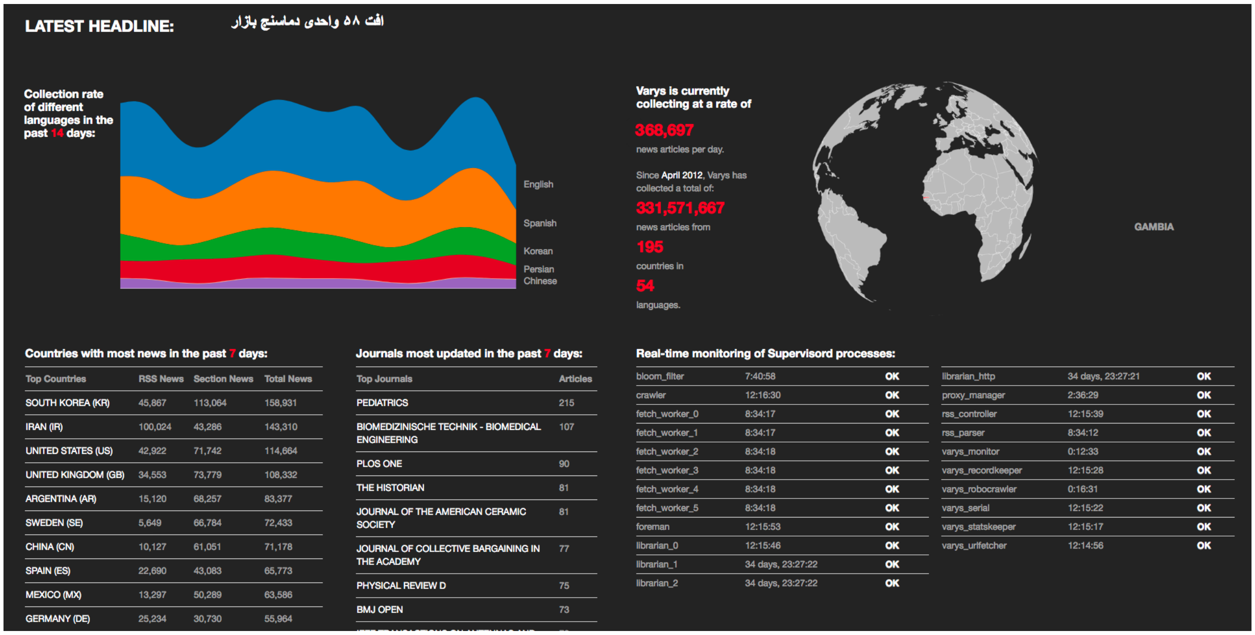
Our master scheduler continuously initiates collection by determining which tasks are due and sending those collection tasks out to the appropriate sub-component. For text documents:

- The scheduler sends a seed webpage to our anonymizing proxy infrastructure, which uses temporary proxies in various geographic locations to request and collect the information. This obscures Kingfisher as the source of these requests and permits us to simulate being local to the content provider to receive localized versions of their content.

- The response is processed for any new URLs in the Hypertext Markup Language (HTML) response, comparing the list of extracted URLs to our list of previously collected URLs.
- Any new URLs are added to the collection list and downloaded through our anonymizing proxy infrastructure.
- The HTML response is parsed through our text extraction tools to eliminate the extraneous information in the response, leaving a clean text document that reads like a newspaper article, and packages the document text and associated metadata into JavaScript Object Notation (JSON).
- The JSON is sent to our Librarian cluster for batch processing and distribution.
- Our Librarian infrastructure is built on a Mongo database cluster capable of handling tens of millions of new data files per day, which exceeds the global OSINT text and multi-media production rate.

The collection runs on a variable cycle, so that news sites are polled for changes about every 15 minutes and periodic journals are polled in line with their print schedule. The timing reflects the rate at which the site adds new content and moves older content to its archives. The timing is a configurable parameter that can be shortened to minimize latency. Our collection infrastructure is fully virtualized and runs with N+1 redundancy.

Our current data dashboard, show below in **Figure 1**, provides an easily ingested view of our current collection efforts. This dashboard was built for our internal purposes and can be modified to better suit the needs of the end user. We can also store multi-media documents, such as images and video clips, and tag them for separate processing while still retaining their original metadata.



LATEST HEADLINE: افت ۵۸ واحدی دماسنج بازار

Varys is currently collecting at a rate of **368,697** news articles per day.

Since **April 2012**, Varys has collected a total of: **331,571,667** news articles from **195** countries in **54** languages.

Countries with most news in the past 7 days:

Top Countries	RSS News	Section News	Total News
SOUTH KOREA (KR)	45,867	113,064	158,931
IRAN (IR)	100,024	43,286	143,310
UNITED STATES (US)	42,922	71,742	114,664
UNITED KINGDOM (GB)	34,553	73,779	108,332
ARGENTINA (AR)	15,120	68,257	83,377
SWEDEN (SE)	5,649	66,784	72,433
CHINA (CN)	10,127	61,051	71,178
SPAIN (ES)	22,690	43,083	65,773
MEXICO (MX)	13,297	50,289	63,586
GERMANY (DE)	25,234	30,730	55,964

Journals most updated in the past 7 days:

Top Journals	Articles
PEDIATRICS	215
BIOMEDIZINISCHE TECHNIK - BIOMEDICAL ENGINEERING	107
PLOS ONE	90
THE HISTORIAN	81
JOURNAL OF THE AMERICAN CERAMIC SOCIETY	81
JOURNAL OF COLLECTIVE BARGAINING IN THE ACADEMY	77
PHYSICAL REVIEW D	75
BMJ OPEN	73

Real-time monitoring of Supervisord processes:

bloom_filter	7:40:58	OK	librarian_http	34 days, 23:27:21	OK
crawler	12:16:30	OK	proxy_manager	2:36:29	OK
fetch_worker_0	8:34:17	OK	rss_controller	12:15:39	OK
fetch_worker_1	8:34:17	OK	rss_parser	8:34:12	OK
fetch_worker_2	8:34:18	OK	varys_monitor	0:12:33	OK
fetch_worker_3	8:34:18	OK	varys_recordkeeper	12:15:28	OK
fetch_worker_4	8:34:18	OK	varys_robocrawler	0:16:31	OK
fetch_worker_5	8:34:18	OK	varys_serial	12:15:22	OK
foreman	12:15:53	OK	varys_statskeeper	12:15:17	OK
librarian_0	12:15:46	OK	varys_urlfetcher	12:14:56	OK
librarian_1	34 days, 23:27:22	OK			
librarian_2	34 days, 23:27:22	OK			

Figure 1: Kingfisher internal data dashboard with geolocation

Multi-Lingual Capabilities

Kingfisher's VARYSS currently collects OSINT in over 90 languages and processes each document in its native format. Our NLP tools vary the sophistication of the processing in each language based on our ability to accurately parse that language, and are limited only by the current state-of-the-art for automatic processing of text in each target language. Kingfisher has extended the full power of VARYSS to non-English languages, with a current focus for enhanced processing in Korean and Persian.

Sample Data Clustering

Grouping, threading, and filtering news is essential to exploit global OSINT when there are hundreds of thousands of articles per day. As a sample demonstration of our news clustering capabilities, VARYSS extracted complex events from news about Poland from July of 2018, shown in **Figure 2** below. We used our common pipeline to cluster the articles and then thread the clusters to identify complex events. This task was accomplished without human input in the document selection or threading processes, validating our ability to perform identification, processing, and relationship mapping of complex news clusters from a sparse news environment.

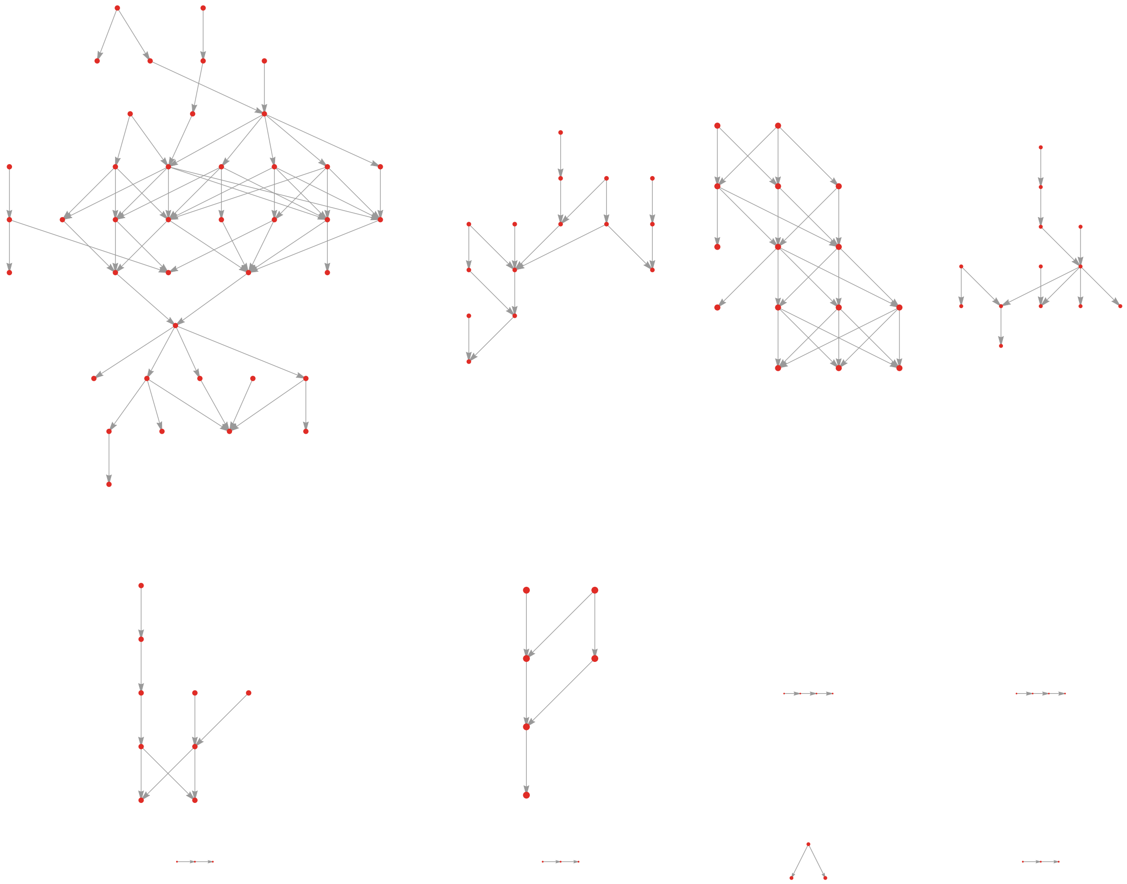


Figure 2: Complex Political Event Interactions Over Time

Even with the sparse English-language news coverage of Poland, we observed six events that occur over multiple days and have branching or merging storylines. We plotted the evolution of the 12 complex events that occurred over at least three days. In the plot, each level is a day, and arrows indicate relationships from a story cluster at time *t* to a story cluster at time *t+1*.

The largest and most complex cluster involves several interrelated stories: (1) a disagreement within the German government over migration; (2) a diplomatic spat between Poland and Israel; (3) judicial reforms within Poland that were designed to entrench the power of the Duda government; (4) the EU response to the actions of the Duda government; and (5) mass public protests over the government's actions.

Sample Knowledge Signaling

Quantitative analysis, particularly regarding the activity of elites, enables understanding of the geopolitics of what DoD has termed Gray Zone conflicts and provides a means to anticipate and develop response options to those types of conflicts. As a sample demonstration of our knowledge signaling capabilities, VARYSS looked at all public statements relating to Wikileaks from hundreds of public figures, over the course of 24 months, including the period before Wikileaks was in the news for publicizing secrets related to the U.S. Government. This information is represented in a 4-axis graph below at **Figure 3**. The dots represent the timing of unlikely comments on this subject area by each named official. The curved graph lines represent the general level of elite discourse on this topic, showing timing and volume. The timing of the individual comments indicate their level of knowledge on the subject, and the relative placement of those dots to the curves of the general level of discussion of the topic show whether they were likely to have been briefed in advance of upcoming disclosures or were reacting to the disclosure. Behavior like this also signals the validity and relevance of a potential leak.

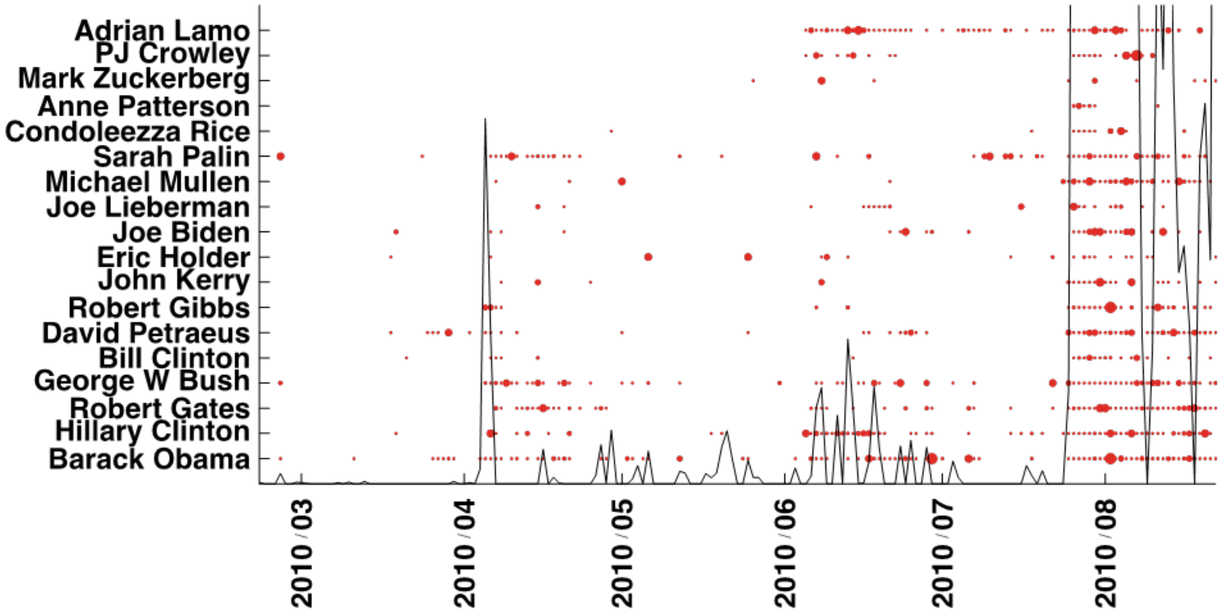


Figure 3: Wikileaks Analysis