MITRE

SERIES
Number 15

INTELLIGENCE AFTER NEXT

THE PREVAILING NARRATIVES ABOUT OPEN-SOURCE INTELLIGENCE ARE MISGUIDED

by Craig Dudley and Cheryl Clopper

The OSINT False Start

The SABLE SPEAR project was the subject of a 2021 article describing an applied Artificial Intelligence (AI) methodology employed entirely within the unclassified domain to understand the global flows of illicit fentanyl.[1] The experiment, largely a punitive journey, worked through implementing this new methodological approach to intelligence analysis. What the team discovered was a fundamental conflict between how the Intelligence Community (IC) was framing the open-source domain, as a *collection space*, and what we began experiencing as the real and unique value of open source, *resolving the "what is happening"* part of the intelligence analysis process. The distinction between the two is significant and illuminates the need to fundamentally change how and where concepts are developed and the way we approach producing timely and comprehensive insight for intelligence customers. This begins with rethinking where and how concepts are built and adjusting the business processes and tradecraft to accommodate these changes.

Convention Rules a Newly Relevant Domain

It is not surprising the IC looked at the open-source domain through the conventions of their well-established and codified business processes.[2] Those conventions place all-source analysts as the central builders of concepts and the creators of finished intelligence – responsible for resolving concepts through navigating an information environment to determine what is known, what is unknown, and assessing what it means. Perhaps unsurprisingly then, analysts spend a significant majority of their time resolving the "what is happening" part of concept development by iterating within familiar information environments to find behaviors unique to the concept and the associations to entities exhibiting those behaviors. A unique characteristic of intelligence, compared to academic discovery, is that analysts have access to a range of sensitive collection resources and, with that, hold a responsibility to develop "collection requirements" as entry to those systems. Analysts interact with collection managers and collection disciplines, like Human Intelligence (HUMINT), Signals Intelligence (SIGINT), and Geospatial Intelligence (GEOINT), to systematically gain greater understanding of "what is happening," so they are better able to characterize the "so what" part of the intelligence process using interpretation and judgment and suggesting opportunities for potential interventions.

The collection management process requires that analysts refine their unknowns to detectable characteristics that are distinctively identifiable in an information environment. These unknowns, submitted as formal collection requirements, are distributed

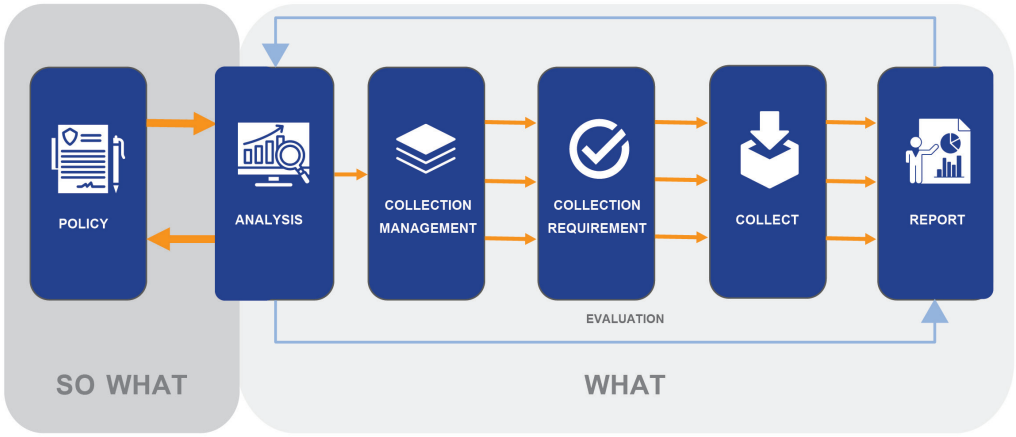

Figure 1: How we build the "what" and "so what"

across the various collection disciplines. In the case of information likely resident in the open domain, they are given to experts trained as Open-Source Intelligence (OSINT) collectors. OSINT collectors use their unique tradecraft to find what the analysts need to enhance their understanding of the "what." Analysts do not ask OSINT collectors to resolve the concept – they ask them to extract from the open domain the specific entities and behaviors that allow them to build the concept. Our current business processes – often referred to as an intelligence cycle, is roughly depicted in Figure 1 with much of the process focused on resolving what is happening.

Over the last decade, analysts have realized that a significant majority of analytically relevant data resides in the open. They need efficient mechanisms to have it brought to them to determine relevance and accuracy before incorporating it into all-source intelligence. Adding to the complexity of the open domain is the prevalence of United States Persons Information (USPI) that requires a special type of protection. While analysts explore the open domain on their own or ask open-source collection professionals to do it on their behalf, information is navigated not just for relevance and accuracy but also for its necessary special handling if determined to fall within the rules that govern USPI.

Through this lens, it was logical that this increasingly relevant and complex domain – the open-source domain – be conceptualized as a collection space that required specialization, trained experts, and unique rules. The creation of the method also followed the pattern of other collection disciplines; we placed it doctrinally as the newcomer to the collection business along with longstanding disciplines such as HUMINT. With that came an expectation that the new collection discipline would bring to analysts entities and behaviors that fit within the concept being explored.

However, the volume of data in the open domain inextricably links it to the technology needed to automatically build concepts by extracting what is relevant, resolving what is true, and sequestering what is special. The unending increase in both users and uses of digital connections[3] puts an undue burden on compute and storage resources, but more importantly, analysts who are overwhelmed by noisy and contradictory data that is simultaneously sparse in many ways.

THE "COLLECT THEN ANALYZE" APPROACH TO USING OSINT IS UNSUSTAINABLE FOR THE IC

Applying Algorithms to Build Concepts based on a Common Ontology

Growing parallel to the rapid expansion of data in the open domain is the advancement of algorithms to find meaningful associations among those data in patterns that can characterize the empirical aspects of unique phenomena. We see this every day in the commercial space, where your phone camera can be pointed at a pair of shoes and those shoes identified as a specific brand which can expand to illuminate where you can purchase those shoes and for what price. Algorithms can separate that "signal from the noise" – resolving what is relevant, true, and special – and isolate signals that, in association, define a concept.[4] They can identify trends worthy of additional attention, illuminate correlations across data too broad or deep for a human to reasonably explore, or accurately perform tedious tasks across a multitude of inputs.

For the IC, algorithms can be used to evaluate a piece of data the moment it is touched within the information environment, immediately resolving questions of relevance (as defined by subject matter experts), truthfulness (accuracy), and whether the signal is categorically "special" (correlated to association with a US person, as an example). Ontologies that uniquely define a concept can be developed, agreed upon, and trained into complex models for the extraction of the behaviors unique to specific intelligence problems and

associations to the entities displaying those behaviors.

Models can be built to resolve each of those concepts at a rapid pace across massive datasets, allowing AI to outpace the human mind in finding associations that define the empirical "what" – even while it cannot, yet, put those renderings into the context of the world we live in, "so what." This analytically relevant data and suite of algorithms can form a Dynamic Foundational Data Fabric (DFDF) specific to an intelligence problem. The DFDF can continuously extract what is relevant, resolve what is true and special, and build the concept that is central to the intelligence problem being explored.

Distinguishing the "what" and the "so what"

A significant advantage of applying an AI methodology is that it introduces a framework for quantitatively evaluating the relative value of datasets. As analysts validate the outputs of the algorithmic work – verifying what is relevant, true, and special – the algorithms will inherently illuminate the datasets that carry a disproportional amount of signal informing those outputs. Under current business processes, it is largely the determination of the analysts for what datasets are, or could be, of value to resolving the "what" part of an intelligence problem.

For intelligence agencies to fully implement the power of AI with intelligence, it is necessary to distinguish between the "what" and the "so what" portions of an intelligence process; the "what" being the empirical aspects of the concept we are trying to understand – whether that be the position of an enemy unit, the person who is a foreign intelligence officer, or an advertisement for illicit fentanyl. The "so what" is very much in the speculative space – informed by a degree of understanding of the "what." Appreciating this distinction will allow us to break from convention and avoid some of the primitive rules that help guide us through the cognitive biases central to tradecraft (there are others) to create the DFDF.

WE MUST RECOGNIZE, FUNDAMENTALLY, THAT THE POWER OF THE OPEN-SOURCE DOMAIN IS APPLYING ALGORITHMS TO DATA WHERE IT RESIDES FOR CONCEPT BUILDING, NOT IN PROCESSING AND DISSEMINATING DATA-POINTS TO ANALYSTS FOR THEIR UNDERTAKING OF CONCEPT BUILDING

The figure below depicts the new methodology where all-source analysts educate the creation of an AI ecosystem (DFDF) to build concepts and illuminate the "what" and lead in validating the outputs as relevant, true, and special. Analysts will always play this integral role as the complexities of the intelligence problem being explored continue to evolve.

As eloquently stated by the Honorable Sue Gordon, former Principal Deputy Director of National Intelligence,

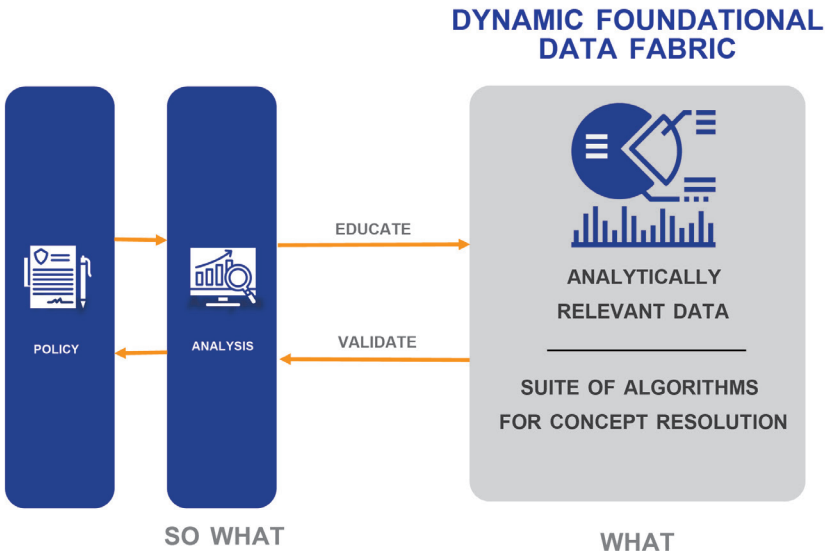

Figure 2: Proposed Business Process

"We have been drafting off the work of our predecessors for so long that we've spent a lot of our years figuring out how we can use the data that already exists, rather than remembering that we are creators too and creating the demand signal for the data we need."[5] The role of the analyst must shift to continuously educate the models, particularly on isolating relevant data and associations that define the phenomena of interest, and then validate those outputs. This is fundamental responsibility that will never end. Performance objectives of analysts must place this activity front-and-center where analysts communicate with data scientists on what they want to see resolved within an information environment and then validate the outputs to confirm relevant, truthfulness, and specialty sufficient to translate into the confident production of finished intelligence.

Trying to apply an "all source" definition[6] and all source analytic tradecraft to resolving the empirical – build concepts – is at once a bit awkward. While estimations about "what could be" face the consequences of cognitive limitations, resolving the empirical – "the what" – is a distinctively different undertaking and an area where commercial applications of AI have proven powerful.

Anchoring Concept Resolution – the "What" – to the Open Domain

Creating a DFDF that consists of analytically relevant data and a suite of ontologically-informed algorithms necessary to resolve concepts must begin and be anchored in the open domain for three reasons:

- A significant majority of analytically relevant data, for any intelligence problem, reside in the open-source domain. With enough effort, the "what is happening" can be pieced together from sensors resident in our surroundings, evidenced in the aftermath of the Boston bombing[7] and the high-profile work conducted by Bellingcat in investigating war crimes in Ukraine.
- The dynamic nature of algorithms and the relationships among them also reside in the unclassified domain. Commercial applications of AI have proven that algorithms can extract concepts from an information environment and render those outputs in both link analysis charts or knowledge graphs, and geospatially for more timely consumption by analysts and investigators. An expectation to move a dynamic data environment and the algorithms needed to separate and associate signals from noise to a controlled domain is unsustainable. Bureaucracy of controlled domains inevitably thwarts the rapid iteration between analysts and data scientists to identify and grow the corpus of signals.
- The sharing space of the unclassified domain is nearly unlimited. Some scholars argue that "data is the new oil" and, if that is the case, having some currency with partners (knowledge of "what is happening") is helpful. Intelligence officers spend much of their time researching "what is happening" within their assigned intelligence problems but the "what" is necessary well beyond intelligence. As trust in public institutions declines,[8] the "what" is needed to build capability, reliability, and transparency. It is needed in the policy domain across multiple agencies to inform opportunities to intervene. The greater clarity of the "what" resolved within the open-source domain means that policymakers could reasonably be informed "right now" in a method that is immediately sharable with any partners of choice.

Building and revealing concepts in the open domain also allows for traditional and non-traditional partners, including academia, to help define the concept. By jointly leveraging the technologies necessary to extract concepts from big-data environments and resolve what is empirically accurate, it broadens the integrity of a collective intervention approach.

While the IC continues to grapple with understanding and codifying the extent to which unclassified data, in sufficient association, must become classified, it is clear that greatest consideration should be given to the insights remaining unclassified.[9] Not least of which is the expectation that our adversaries are able to

resolve the same discoveries within what is a common competitive space for discovery.As the IC considers the need for classification guidelines, the standard should be to classify unclassified insights when it becomes 'overwhelmingly clear' that the insights and methodology for deriving them would pose a risk to national security, if exposed. But to be sure, the only way to know what is secret is through a commanding understanding of what is not.

A Clean Start for OSINT

First and foremost, the IC must embrace the big data environment of the open domain as powerful in building the ecosystems necessary for automated concept development – to isolate what is relevant and truthful from what is not and identify what is categorically special and worthy of a unique type of protection. This power is not singularly within the richness of the data environment itself, but in the simultaneous employment of complex algorithmic environments that do the resolutions the moment data are touched within that dynamic environment – the DFDF.

For the IC to make progress toward the DFDF, we must distinguish between resolving the "what is happening" from the "so what" and accept that the journey to do both is strikingly different than the way we think about and do things now. This new space will require new specializations, trained experts, and rules that govern how this technology is safely and effectively protected and integrated. We should aspire to a future where:

- Analytically relevant data and applied algorithms for concept resolution (the DFDF) in support of the "what" resides on the open domain
- Analysts shift from looking for each "what" to refining characteristics to define the concept in partnership with data scientists
- The IC expands its partners for concept development and reliably illuminates the concept's existence to inform the collective intervention space

The IC can get off the starting blocks in the AI race by rethinking where and how concepts are built – surrendering the initial work in concept development to algorithms in the open-source domain. This journey would gradually shift the work of the "what is happening" to data science and correspondingly free up analysts' time to think and write about what these activities mean and what opportunities exist for intervention.

Notes

1. Dudley, Craig A. 2021. "Lessons from SABLE SPEAR: The Application of an Artificial Intelligence Methodology in the Business of Intelligence." Studies in Intelligence 65 (1): 7-14. https://www.cia.gov/resources/csi/studies-in-intelligence/volume-65-no-1-march-2021/lessons-from-sable-spear-the-application-of-an-artificial-intelligence-methodology-in-the-business-of-intelligence/
2. Office of the Director of National Intelligence. 2015. "Intelligence Community Directive 203: Analytic Standards." 01 02. Accessed 09 20, 2022. https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf
3. Kemp, Simon. 2022. Digital 2022: Global Overview Report. DataReportal. 01 26. Accessed 09 20, 2022. https://datareportal.com/reports/digital-2022-global-overview-report
4. Pearl, Judea, and Dana MacKenzie. 2018. The Book of Why. New York: Basic Books.
5. Myatt, Summer. 2022. Honorable Sue Gordon Outlines Intelligence Community's Top 5 Data Problems. 10 12. https://www.govconwire.com/2022/10/honorable-sue-gordon-outlines-intelligence-communitys-top-5-data-problems/
6. National Institute of Standards and Technology. n.d. all-source intelligence. Computer Security Resource Center. Accessed 09 21, 2022. https://csrc.nist.gov/glossary/term/all_source_intelligence NIST defines all-source intelligence as "Intelligence products and/or organizations and activities that incorporate all sources of information, most frequently human resources intelligence, imagery intelligence, measurement and signature intelligence, signals intelligence, and open source data in the production of finished intelligence"
7. National Center for Audio & Video Forensics. 2021. Surveillance and Solving the Boston Bombing. Accessed 10 10, 2022. https://ncavf.com/press/surveillance-and-solving-the-boston-bombing/
8. Pew Research Center. 2022. Public Trust in Government: 1958-2022. 06 06. Accessed 09 20, 2022. https://www.pewresearch.org/politics/2022/06/06/public-trust-in-government-1958-2022/
9. Harris, Shane, Karen DeYoung, Isabelle Khurshudyan, Ashley Parker, and Liz Sly. 2022. Road to war: U.S. struggled to convince allies, and Zelensky, of risk of invasion. The Washington Post. 08 16. Accessed 10 10, 2022. https://www.washingtonpost.com/national-security/interactive/2022/ukraine-road-to-war/

.

Authors

Craig A. Dudley is a division chief in the US Defense Intelligence Agency. During his 20-year career with DIA, most of it overseas, he has had experience in capacity building, collection management, and all-source analysis, and served multiple tours at combatant commands and the Joint Staff. He holds a Doctorate in Comparative Intelligence from the University of Botswana, where his research focused on applied intelligence models.

Cheryl Clopper serves as the Analysis Division Chief Engineer at The MITRE Corporation, leading and guiding high impact technical work. She accelerates cloud computing implementation for mission needs with software prototypes and applied systems engineering across the Intelligence Community and the military's combatant commands.

Intelligence After Next

MITRE strives to stimulate thought, dialogue, and action for national security leaders developing the plans, policy, and programs to guide the nation. This series of original papers is focused on the issues, policies, capabilities, and concerns of the Intelligence Community's workforce as it prepares for the future. Our intent is to share our unique insights and perspectives surrounding a significant national security concern, a persistent or emerging threat, or to detail the integrated solutions and enabling technologies needed to ensure the success of the Intelligence Community.

About MITRE

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

About DIA

DIA's diverse workforce provides intelligence on foreign militaries to prevent and decisively win wars. We support warfighters, defense policymakers, and force planners in the Department of Defense and the Intelligence Community. We plan, manage, and execute intelligence operations during peacetime, crisis, and war.

MITRE

FEBRUARY 2023