DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

ORNL/LTR-2024/3674

AI for Nuclear Safeguards Verification

Francisco Parada
Nathan Martindale
Alisa Reasor
Scott Stewart
Lindsey Ukishima

October 2024

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/LTR-XXXX/XXX

Nuclear Nonproliferation Division

AI FOR NUCLEAR SAFEGUARDS VERIFICATION

Francisco Parada
Nathan Martindale
Alisa Reasor
Scott Stewart
Lindsey Ukishima

October 2024

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

ABSTRACT

The International Atomic Energy Agency (IAEA) utilizes AI/ML to analyze open-source information to verify the completeness of State declarations of nuclear activities. AI/ML is used to help automate processes and analysis of large datasets, including satellite imagery and unstructured data from publications, to improve efficiency and effectiveness of safeguards implementation. There are several considerations when proposing the use of AI/ML in nuclear safeguards, including the need for large, unbiased datasets and the risk of AI-generated fake information. Two of the more achievable, near-term uses of AI/ML in IAEA safeguards include assisting with processing satellite imagery and automating knowledge management and extraction.

1. INTRODUCTION

As identified in the *Development and Implementation Support Programme for Nuclear Verification 2024–2025,* the International Atomic Energy Agency (IAEA) is currently “leveraging AI breakthroughs, including generative AI, for enhanced ML projects in safeguards information analysis, ensuring responsible development, fostering knowledge sharing with external experts, and incorporating subject matter expertise into data-driven systems” [1]. Moreover, AI, with its capacity for automation and data processing, supports many of the underlying objectives identified in the IAEA top priority capabilities for 2024–2025 [2].

The collection and evaluation of all safeguards-relevant information is part of the process for IAEA safeguards implementation. Three sources are considered: information provided by the state, information obtained from IAEA safeguards activities, and all other safeguards-relevant information [3]. Open-source information, including but not limited to scientific and technical publications, news articles, government records, trade databases, and social media, is crucial for detecting possible undeclared nuclear activities. This information is largely unstructured data in various formats, including print, electronic databases, web pages, audio, and video. Open-source information undergoes a validation process by subject matter experts before being integrated with other safeguards-relevant information. This comprehensive approach of combining open-source information with data collected from in-field activities strengthens the assessment of state-provided information, particularly regarding the accuracy and completeness of state declarations [4],[5].

The anticipated deployment of advanced reactors will significantly increase the volume of information, facilities, and material that will require safeguards. This presents both a need and an opportunity for AI to enhance productivity across various safeguards activities because it can reduce the gap between available analyst resources and the exponential growth of potentially relevant information. AI/ML initiatives for safeguards should aim to support this need by enabling the IAEA to efficiently process, identify, and utilize safeguards-relevant information while optimizing resource allocation.

Employing AI/ML models and algorithms in parallel with expert analysis during the validation process can offer valuable alternative assessments, ultimately enhancing safeguards decision-making. Additionally, AI can significantly streamline time consuming processes that involve extensive manual labor, such as analyzing the growing volume of data generated by IAEA video surveillance, other safeguards equipment, and declarations and reports provided by the state itself.

Although AI/ML offers the potential to enhance safeguards at different stages and activities—particularly in applications that include collection and analysis of unstructured data—its application continues to have challenges. AI/ML models require large amounts of data to learn from, and generating relevant datasets for safeguards can be difficult. One challenge in utilizing AI for safeguards is the potential for bias within the datasets used to train the algorithms and language models. This has a direct implication for safeguards because the IAEA must ensure that its safeguards methods are technically sound and unbiased when reaching safeguards conclusions [6].

2. SATELLITE IMAGERY AND AI/ML

The IAEA has a 20-year history of utilizing open-source satellite imagery analysis to support its safeguards conclusions and improve the effectiveness and efficiency of safeguards implementation. In 2023 alone, the agency acquired 1,768 satellite images [7]. This satellite imagery aids in supporting inspections and reviewing the completeness of state declaration information by enabling the identification of facility modification for comparison with the declared information. This process, known as *change detection*, involves identifying and measuring how an object or phenomenon changes over time. By comparing satellite images taken at two different times, each pixel or object in the first image is compared to its counterpart in the second image to determine the extent of change between the two moments. Other applications of satellite imagery include verifying declared site layouts, detecting undeclared activities (e.g., new mining or construction), monitoring decommissioned facilities, aiding inspection planning and reporting, and tracking the operational status of facilities [8],[9].

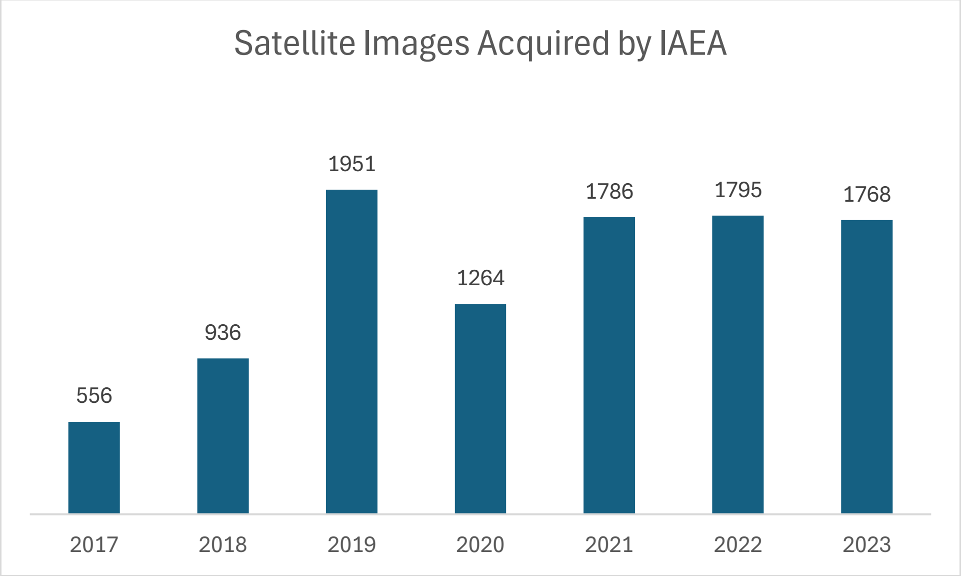

Figure 1. Number of satellite images acquired by the IAEA by year [10].

The increasing availability of satellites, sensors, and imagery has led to a significant rise in the frequency of observations (Figure 1). This in turn, places limitations on the types of activities that can be conducted without detection and enhances the reliability of analyses that involve tracking ground activity and identifying trends [11]. Continuous access to satellite imagery with different constellations provides the IAEA with opportunity to better monitor facilities of interest and increases the probability of detection of attempts identify confirm the activities in the State declaration and detect the appearance or removal of relevant objects. The different sensor technologies available in commercial satellite imagery also provides the IAEA with the ability to monitor outputs from facilities such as vapor plume emissions from cooling towers or steam discharges into surrounding water that can be used to estimate the nuclear material use or production in a facility.

Satellite imagery provided by third-party organizations can also assist the IAEA in reaching safeguards conclusions. Indeed, cases involving undeclared nuclear activities investigated by the IAEA have originated from such third-party information. This is permitted under the IAEA Statute, Article VIII.A: "Each member should make available to the Agency such information as would, in the judgment of the member, be helpful to the Agency" [12]. However, the IAEA will not take this information at face value and conducts a thorough assessment, depending on the type of information and data provided this could range from using safeguards relevant data to "google searches and common sense" [13] to verify that such information provided by a third party is legitimate and of safeguards relevance.

As identified by the IAEA, the increasing sophistication of AI content and manipulation techniques poses a significant challenge. "For example, while generative AI offers new opportunities for Safeguards, it also poses the risk of misuse through the creation of fake information in open sources. There is a demand to reliably incorporate expert knowledge into data-driven systems, and there is a need for common guidelines and validation procedures to address the risks posed by new technologies" [1]. States could potentially use AI to create forged or spoofed satellite imagery, either to falsely implicate other states or to conceal their own illicit activities.

The deliberate alteration of satellite imagery to conceal or misrepresent ground features is not a new concern. An illustrative case is a Swedish state agency that obscured the headquarters of a government organization on aerial photographs by superimposing trees and fields over the buildings. This practice underscores the broader issue of image manipulation and highlights the relevance of AI in verifying the authenticity of satellite images [14]. AI can also be used to detect anomalies because it can be trained to recognize inconsistencies or indications of manipulation, thereby aiding the IAEA in ensuring the integrity of information used for critical safeguards decisions. AI can also provide an independent assessment, thereby serving as a secondary evaluation to complement expert analysis from a safeguard's perspective.

Despite the general challenge of limited relevant data for safeguards, the volume of satellite imagery continues to increase, and models can benefit from a wide variety of datasets and commercially available imagery. However, nonproliferation-specific images may be limited due to sensitivity concerns, few historical examples, and the variety of pathways that actors might pursue. Furthermore, converting satellite imagery into useful safeguards information remains a challenge due to the lack of labeled training data and the complex and time-consuming process of creating labeled data. It requires expert knowledge and careful annotations that the model can utilize to differentiate objects or activities of interest from the background.

3. KNOWLEDGE MANAGEMENT AND EXTRACTION

Knowledge management is a critical component of open-source analysis. Knowledge management is a responsibility shared across several groups within the IAEA, but in particular the State Factor Analysis section (ISF)[1] within the Division of Information Management maintains an open-source library of safeguards-relevant, open-source information [15]. This library must be regularly updated with new information and documents that they collect and evaluate daily. The library is updated with a tool called the Open-Source Information System (OSIS) [16]. OSIS continuously crawls, scrapes, and collects information from the web and provides the analyst with an interface from which they can review, tag, and annotate collected information. A diagram of this system is shown in Figure 2. Knowledge management and extraction is relevant to most steps in this diagram, and many of them can be further explored and improved through the application of knowledge management principles and AI/ML. Examples include more targeted/filtered feeds for analyst review through better relevancy prediction, automatic

[1] "The State Factor Analysis Section (ISF) is responsible for the collection, validation, evaluation, analysis, and dissemination of Safeguards-relevant open source information for the Department. This happens daily as an essential component of performing continuous State Evaluation and supporting Safeguards implementation." [1]

classification, organization, and extraction of documents and associated metadata as well as mechanisms and tools for querying the open-source library and ad-hoc construction of structured information views to assist an analyst in evaluating and contextualizing new information.

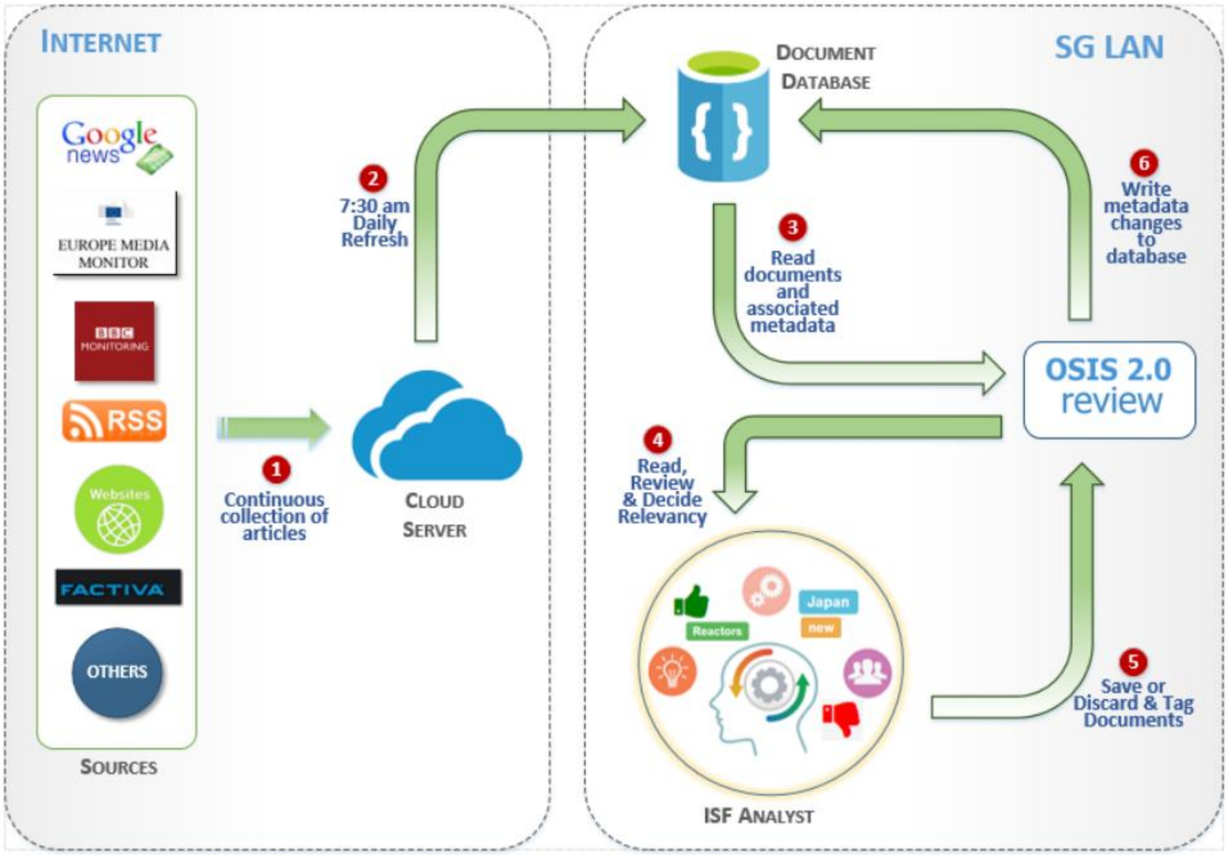

Figure 2. OSIS workflow. Figure from Skoeld, Courbon, and Spence [16].

Open-source analysts at the IAEA employ search queries to locate relevant scientific publications for evaluating the consistency between a state's declared nuclear activities and actual nuclear activities. Although an analyst's experience plays an important role in the performance of these searches to pinpoint relevant information for collection and analysis, the searches are still fundamentally constrained by the specific presence of each term. Figure 3 highlights the need for more effective approaches to finding and filtering relevant sources given the quantity of open-source information available. The IAEA previously prototyped OSIS improvements that included the use of AI/ML techniques to predict and rank relevance of new documents as well as a physical model classifier that categorizes documents into aspects of the fuel cycle to which they are most relevant. That work indicated interest in future efforts to improve capability for similarity measurement, with applications for detecting new research content as well as matching common research across collaborating states.

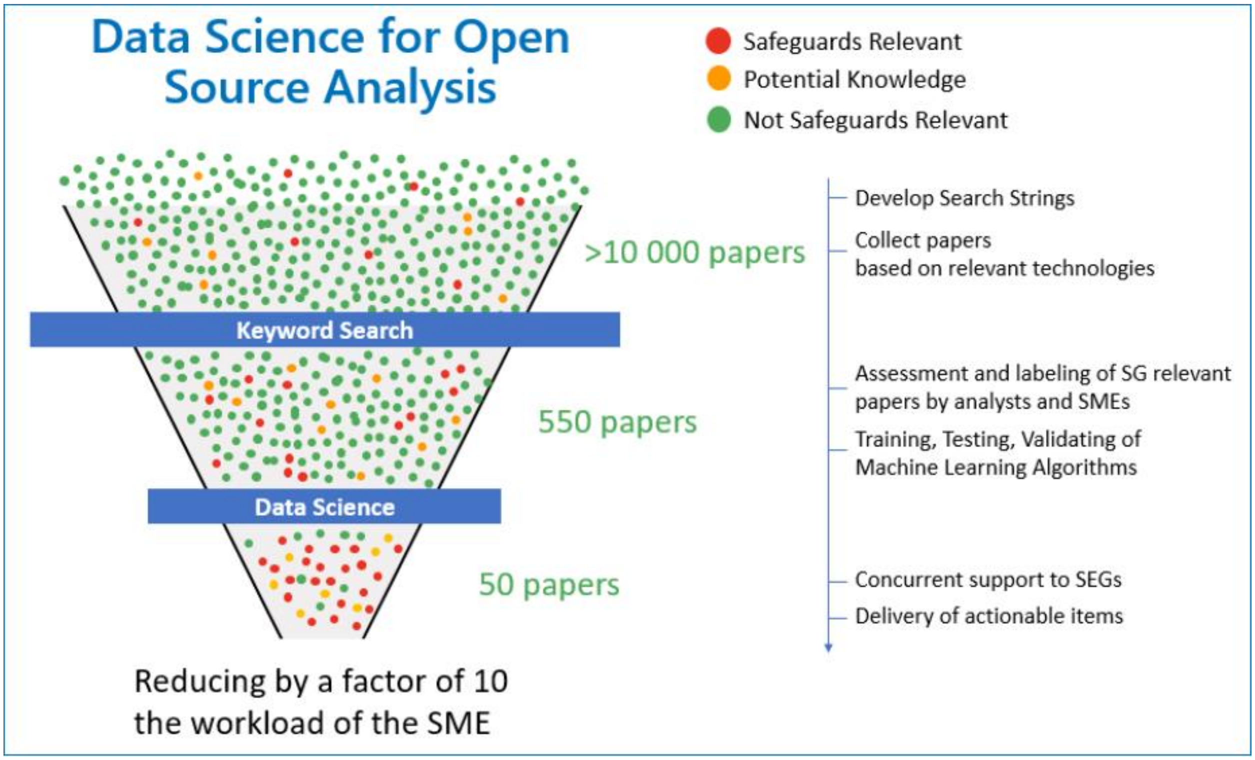
Data Science for Open Source Analysis
● Safeguards Relevant
● Potential Knowledge
● Not Safeguards Relevant
>10 000 papers
Keyword Search
550 papers
Data Science
50 papers
Develop Search Strings
Collect papers based on relevant technologies
Assessment and labeling of SG relevant papers by analysts and SMEs
Training, Testing, Validating of Machine Learning Algorithms
Concurrent support to SEGs
Delivery of actionable items
Reducing by a factor of 10 the workload of the SME

Figure 3. Open-source data analysis workflow [1].

Modern developments in the field of natural language processing (e.g., large language models [LLMs]) are of interest to the IAEA and could enhance analyst workflows in knowledge management. One issue with using an LLM for a given application is that the LLM can use only the information it was trained on. So, if more, new, or different information is needed for a given use case, then the LLM will need to be retrained on that information or have the user include a significant amount of that information in the conversational prompt when using the model. Additionally, LLMs can provide incorrect information in response to a question, a phenomenon commonly known as *LLM hallucination*.

Retrieval augmented generation (RAG) [17] is one approach that can at least partially mitigate these issues. In a RAG-based system, a vector knowledge database is created by embedding a collection of source documents. At generation time, relevant snippets from the vector database are retrieved and injected into the prompt. Construction of a vector database is substantially less compute intensive than the process of fine-tuning an LLM and provides a level of visibility and interpretability for what the LLM is using to create a response. Additionally, information in the database can be updated, removed, or dynamically constructed as needed, such as when an open-source library is modified.

The utility of a RAG-based system is based on the quality, organization, and method of retrieval from the database, and knowledge management techniques may improve this process. Atomic snippets of information that are heavily interlinked form a network, and this coarser form of a knowledge graph could allow more intelligent context retrieval based on network traversal. This same process could also directly support analyst activities such as compiling reports, which involves collecting, ordering, and refining knowledge relevant to a particular report subject. LLMs and RAG-based systems may provide a means for analysts to more effectively follow knowledge management principles in how the open-source library is maintained and updated. LLMs equipped with appropriate prompts may allow automatic extraction of summaries, subsections, and link prediction such that information added into the library fits the atomic and interlinked properties discussed above.

4. REFERENCES

[1] IAEA, *Development and Implementation Support Programme for Nuclear Verification 2024–2025*, STR-405

[2] IAEA, *Top priority capabilities Update for 2024–2025*

[3] IAEA, *IAEA Safeguards Serving Nuclear Non-Proliferation*, 2015

[4] IAEA, *IAEA Safeguards Glossary*, 2022

[5] Schneeweiss, P., T. Stojadinovic, and Z. Abbali et al., “Extracting the Signal From the Noise – The Role of Artificial Intelligence in the Analysis of Safeguards Relevant Information,” 2023.

[6] IAEA, “Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology,” 2022.

[7] IAEA, *IAEA Safeguards in 2023*, 2024

[8] Schneeweiss, P., T. Stojadinovic, and S. Baude, “The IAEA’s Innovative Approach to Address the Challenges in the Collection and Analysis of Safeguards Relevant Information,” 2023.

[9] Niemeyer, I. and S. Nussbaum, “Change Detection: The Potential for Nuclear Safeguards,” R. Avenhaus, N. Kyriakopoulos, M. Richard, and G. Stein (eds), *Verifying Treaty Compliance*, Springer, Berlin, Heidelberg, 2006.

[10] IAEA, IAEA Safeguards infographics in 2023, 2022, 2021, 2020, 2019, 2018, 2017.

[11] Moric, I., “Capabilities of Commercial Satellite Earth Observation Systems and Applications for Nuclear Verification and Monitoring,” *Science & Global Security*, 30(1) (2022): 22–49.

[12] IAEA, IAEA Statute, 1956

[13] ElBaradei, M., The Age of Deception, Nuclear Diplomacy in Treacherous Times, Metropolitan Books, 2011.

[14] Geens, S., “Sweden plays hide and seek with maps,” https://ogleearth.com/2006/04/sweden-plays-hide-and-seek-with-maps/.

[15] IAEA, *Development and Implementation Support Programme for Nuclear Verification 2022–2023*.

[16] Skoeld, T., F. Courbon, and K. Spence, “OSIS 2.0: Optimizing Analyst-Driven Automation of Open Source Information Collection and Processing for Safeguards State Evaluation.”

[17] Piktus, A., F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks,” NeurIPS 2020.