Computational Methods for Physical Model Information Management: Opening the Aperture

Jack Crowley, The MITRE Corporation, jdcrowley@mitre.org
Danny Gagne, Department of Safeguards, IAEA D.Gagne@iaea.org
Dan Calle, Department of Safeguards, IAEA Dan.Calle@iaea.org
John Murray, Department of Safeguards, IAEA j.murray@iaea.org
Remzi Kirkgoeze, Department of Safeguards, IAEA r.kirkgoeze@iaea.org
Frank Moser, Department of Safeguards, IAEA F.Moser@iaea.org

Introduction

The volume, velocity and diversity of data available to analysts are growing exponentially, increasing the demands on analysts to stay abreast of developments in their areas of investigation. In parallel to the growth in data, technologies have been developed to efficiently process, store, and effectively extract information suitable for the development of a knowledge base capable of supporting inferential (decision logic) reasoning over semantic spaces. These technologies and methodologies, in effect, allow for automated discovery and mapping of information to specific steps in the Physical Model (Safeguards’ standard reference of the Nuclear Fuel Cycle).

This paper describes an integrated service under development at the IAEA that utilizes machine learning techniques, computational natural language models, Bayesian methods and semantic/ontological reasoning capabilities to process large volumes of (streaming) information and associate relevant, discovered information to the appropriate process step in the Physical Model. The paper will detail how this capability will consume open source and controlled information sources and be integrated with other capabilities within the analysis environment, and provide the basis for a semantic knowledge base suitable for hosting future mission focused applications.

The Challenge: Growth in Volume, Velocity, and Diversity

A component of the IAEA mission is to verify State compliance with commitments under the Non-Proliferation Treaty and other non-proliferation agreements.1 The additional protocol increases the reporting obligations of many States to provide the IAEA with a fuller and clearer understanding of their nuclear activities. The declarations made under these various agreements are then checked for correctness and completeness. Verifying the completeness of the declarations is a particularly challenging task as it is difficult to prove a negative; as such, the challenge is to perform as comprehensive an analysis as possible over any available sources of information that may be relevant to the declarations.
As a part of the verification process, analysts form evidence-based hypotheses regarding nuclear related activities of the States under the different non-proliferation agreements. Evidence may take many forms: sensor measurements, nuclear material accountability, etc. Increasingly, evidence (both confirming and disconfirming) of nuclear related activities is discovered in the growing web of publicly available information including scientific publications, university and industrial web sites, news streams and social media streams.

The existing volumes of information are staggering. As an illustration, consider:

- An estimated 2.7 Zetabytes\(^{ii}\) of data exist in the digital universe today.\(^{iii}\)
- In September of 2014 there were over 1,000,000,000 registered host names\(^{iv}\) representing unique web sites while in the year 2000 there were approximately 17 million sites. Note that not all websites are active.
- As of January 2014, Google Scholar has indexed over 100 million technical articles and patents.\(^{v}\)

These numbers are increasing exponentially, challenging the analyst’s capacity to identify and interpret the available information.

**IAEA Content Reification (ICore) Service: Natural Language Processing (NLP) and Reasoning**

The IAEA Department of Safeguards has been developing an automated natural language processing (NLP) and a big-data analytic capability prototype to help manage the challenge posed by an exploding digital universe. This effort has the following objectives:

1) Match documents to the appropriate process step within the IAEA process model of the nuclear fuel cycle, known as the IAEA Physical Model This categorization allows the Department to perform analysis on the documents that are relevant to specific technical objectives that are part of the State evaluation process.

2) Match documents to specific States based on geospatial and entity information in the document.

3) Be horizontally scalable to efficiently process large volumes of information.

4) Feed a knowledge repository suitable for analytical inquiry across the corpora of processed documents.

5) Iterate over feedback from the Department to improve results and integration into safeguards’ processes.

ICore’s integrated natural language processing engine and ICore’s latent semantic text categorization framework are central to meeting these efforts. Together, these two complementary capabilities provide a scalable and robust means for addressing the challenge posed by the exponential expansion of available information. Both of these capabilities will be discussed in the following sections of this paper. Natural language processing is a subfield of computer science and linguistics responsible for the development of techniques that allow computers to process and understand human languages.
There are many tasks associated with the field of NLP including document summarization, machine translation, named entity recognition, speech recognition, and sentiment analysis. One of ICore’s principal tasks is Natural Language Understanding. Natural Language Understanding involves identifying the appropriate semantic interpretation of an expression and transforming that interpretation into a computable structure suitable for automated reasoning. This differs from the common NLP task of named entity recognition (NER) where the objective is to associate tokens in a text stream with common classes such as PERSON, ORGANIZATION, and LOCATION.

For example, consider the following: “Dr. Jones from the University of Deep Thought in East Mardina discovered an accelerated centrifugation method.” A natural language processing system focused on named entity recognition will associate no semantic interpretation with the extracted tokens: “PERSON from the ORGANIZATION in LOCATION discovered an accelerated THING.” Semantics is involved with the meaning of the tokens, not just their category. A natural language understanding system such as ICore will associate extracted concepts with a formal knowledge model suitable for performing automated reasoning tasks.

In ICore, formal knowledge models are represented using the web ontology language OWL, a standardized method for formal knowledge representation. ICore is capable of incorporating any knowledge domain represented as an OWL ontology. The core ontology currently integrated with ICore is a representation of the INIS (International Nuclear Information System) knowledge domain containing more than 30,000 concepts.

Using the INIS ontology, ICore is capable of resolving the extracted objects to the appropriate class in the ontology (INIS Centrifugation) and reason that the text has a probability of being associated with an enrichment process. Figure 1 is an illustration of class relationships in the INIS ontology supporting this automated reasoning task.

![Figure 1 ICore Ontology Reasoning Example](image-url)
The ICore service extracts persons, organizations, location data, relative location data, temporal data, relative temporal data, measurement information, and geopolitical entities as well as extracting semantically identifying concepts identified in any ontology under ICore management. All extracted objects are persisted in an online knowledge base, suitable for supporting further analysis. This knowledge base will be discussed later in this paper.

### IAEA Content Reification (ICore) Service: Text Classification to the Physical Model

Text categorization is the activity of labelling natural language texts with thematic categories from a predefined set. It is the intersection between machine learning and information retrieval. Typically generative or discriminative probabilistic models, relying on large collections of labelled (training) data sets are developed to perform the classification tasks. These models rely on the analysis of the statistical distribution of individual words or tokens in a specific document versus the overall distribution of these words in the training corpus to assign a document to its most likely category.

As an alternative to these probabilistic models, ICore uses a reduced semantic vector space approach, Latent Semantic Analysis (LSA), to augment its ontological reasoning capabilities and match documents to the physical model. Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual meaning of words by statistical computations applied to a large corpus of text. The idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of constraints that determines the similarity of meaning of words and sets of words to each other.

The solution uses a training corpus consisting of IAEA’S physical model documentation augmented with other nuclear domain specific documents, and a large collection of deliberately unrelated documents. The corpus is transformed into a large term-by-document matrix. This matrix is then factored into a set of major eigenvectors and eigenvalues using singular value decomposition. The number of eigenvectors represent the reduced space dimension of the original matrix, and are used to provide a similarity measure between unknown documents (documents to be categorized) and documents in the training corpora. Unknown documents are represented as term vectors projected into the lower dimensional eigenspace representing the corpora of known documents. A similarity measure between the reduced, unknown document and the reduced nearby pre-categorized documents in the corpora is calculated (cosine similarity) and the appropriate categorizations are made based on a similarity threshold score.

Unlike the traditional Latent Semantic Analysis technique which considers a document to be an unordered “bag of words,” ICore first processes documents through the natural language processing components and treats multi-term extracted entities as a single term.

For example, using the traditional LSA “bag of words” approach, the sentence “Alice Jones studies pebble bed reactors” would be modeled as a set of six tokens (Alice, bed, Jones, Pebble, reactors, studies). ICore’s method of first processing the text through the NLP pipeline results in the text being modeled as a set of three tokens (Alice_Jones, pebble_bed_reactors, studies). Measures on how this approach
improves the precision and recall of ICore’s ability to categorize documents with respect to the physical model are under development.

ICore’s modular design can support multiple classifiers to enhance matching documents to the physical model.

**Architecture**

ICore has been designed as a high performance component-based service to fit within the IAEA’s information collection, processing, and persistence pipeline. The principal components of the application are as follows:

- The Natural Language Processing System composed of a Conditional Random Field based model training and evaluation framework for extracting entities and concepts from data streams.
- An ontology reasoning system. This system is designed to perform descriptive logic reasoning over extracted concepts and ontologically typed entities.
- A server designed to manage multiple simultaneous clients over a REST interface.
- A multithreaded job scheduler and monitor for asynchronous file based operations.

![Figure 2: ICore Architecture](image)

The system is designed to take advantage of the modern high-core density available on most CPUs to maximize throughput. Initial performance monitoring has shown a throughput of 1 document per core per second with a mean document length of 1000 words. On a four core, eight hyperthreaded core, Intel I7 based server this equates to a single server being able to process over 600,000 documents per day.

To achieve near horizontal scaling across CPU cores, common resources are pooled and made available for all processes. The number of shared resources is configurable by an administrator.
For scalability beyond a single server, processing jobs can be distributed between servers. No synchronization between the servers is required resulting in true horizontal scalability. Job distribution across servers simply requires distribution of files to be processed to shared network file system directories. Each server monitors these directories and will distribute jobs among its available processing cores/threads.

For file system based processing, ICore produces an output XML file which includes meta-information on the file (source, processing time), extracted concepts and entities, and scores in the range of (0,1) reflecting similarity of concepts and entities in the document to different steps in the physical model. For REST based operations, the same XML formatted output is returned to the caller over the requesting socket.

**Data Sources**

Documents processed by ICore are sourced from multiple locations. Currently, International Nuclear Information System (INIS) and European Patent Office (EPO) documents are routed to an ICore-monitored directory for automated processing. The horizontal scalability of the system allows for any number of additional sources to be processed.

The solution under development at the IAEA also includes an open source monitoring capability, the “IAEA Get Open-source Retrieval (IGOR)” engine, and a solution under development for crawling and retrieving documents from specific web sites or domains called IDIG.

The IGOR utility monitors RSS feeds which monitor global news sources for stories related to different physical model processes. Retrieved documents are transformed to text using the Apache TIKAxix framework. This framework detects and extracts metadata and text content from various document formats including HTML, PDF, and common office formats. IGOR routes collected information to ICore for extraction and classification against the Physical Model.

The IDIG utility being constructed around the Apache NUTCHxvi framework is designed to crawl specific web sites or all documents with a sub-domain of specific web sites, process the retrieved data through Tika and route the resulting text documents to ICore for processing. The objective is to facilitate the automated generation of people, organizations and other entities and concepts related to a specific target resource such as a corporation, a specific reactor, or an engineering department at a specific university.

The output of ICore is forwarded for processing by other IAEA services (Palantir, SharePoint, etc.) and also persisted in a data storage environment currently under development for subsequent analysis.

**Knowledge Base Development**

ICore’s initial objective is to categorize documents with respect to the IAEA Physical Model. As discussed in this paper, documents processed by ICore include extracted entities (people, organization, locations etc.) and concepts often associated with other entities. This corpus represents a rich repository capable
of supporting data mining and analysis in an information space spanning time, concepts, geospatial location and entities.

To support this analysis, the prototype includes the development of a hybrid NoSQL virtualized environment, the IAEA Data Environment for Analysis (IDEA), to serve as the data repository for extracted information.

The data environment needs to be capable of storing multiple data formats to support different types of analysis. The initial IDEA environment consists of:

- **MongoDB**\(^{xvii}\), a structured (JSON) document repository for persisting each of the documents processed through ICore.
- **SOLr**\(^{xviii}\), an indexed data store for persisting source documents as they are collected by IGOR and IDIG;
- ** Accumulo**\(^{xix}\), a big-table data store built on Hadoop supporting data mining, time series analysis, and other analytics through map-reduce jobs and domain specific machine learning;
- **Neo4J**\(^{xx}\), an open source graph database suitable for storing extracted entities, the relationships between entities, and attributes of these entities.

This environment, continuously populated by ICore and ingested special collections, is intended to provide an analytical repository for the discovery of trends, patterns and latent relationships between managed entities. In the future it is anticipated that critical information discovered within the IDEA environment will be exported to other IAEA analytical environments such as Palantir to support analyst reasoning and hypothesis.

**Usage Illustration**

The following is an illustration of ICore’s capability. A web utility has been developed to allow engineers to interact directly with ICore for evaluation and debugging purposes. This web utility will be used for this illustration, however ICore and the data collection agents are intended to be run as automated server applications.

IGOR monitors RSS feeds developed by IAEA subject matter experts. Below is a typical document retrieved by IGOR, a recent patent application\(^{xxi}\) that will be used for the sample document.
IGOR extracts the text from the document and forwards to ICore for extraction and text classification. The extracted items and document classification are shown in figure 4 and figure 5 below.
As observed in figure 5, the document is correctly associated with the enrichment process, with the highest score given to Molecular Laser Enrichment.

To further illustrate the robustness of the approach, references to “isotope enrichment” and references to lasers have been removed from the document and the document again processed by ICore.

Items crossed out in the text below represent the modified text to be processed:

“Isotope enrichment by laser activation wherein a multi-isotopic element Q, like Uranium, Silicon, Carbon is incorporated into gaseous QFn, QF6, QF4, QOmFn, etc and diluted in gas G like He, N2, Ar, Xe, SF6 or other inert gas; and wherein that mixture is cooled by adiabatic expansion or other means encouraging formation of dimers QF6:G in a supersonic super-cooled free jet; and wherein that jet is exposed to laser photons at wavelengths that selectively excite predetermined molecules IQF6 to IQF6*, thereby inducing rapid VT conversions and dissociations of IQF6*:G→IQF6+G+kT, while leaving non-excited dimers jQF6:G intact; and wherein a skimmer separates the supersonic free-jet core stream containing heavier IQF6:G dimers from lighter core-escaped IQF6-enriched rim gases. Particularly an advanced technique is disclosed to enrich IU6 by free jet expansion and isotope-selective dimerization suppression, utilizing a molecular CO laser and intra-cavity UF6 irradiation with laser lines overlapping predetermined IU6 absorptions; and providing multiple free jet separator units irradiated by one laser beam, thereby enhancing process economics.”

The resulting classification produced by ICore are shown below in figure 6 and illustrates the ability of the classification system to make reasonable assignment of documents to the appropriate physical model step.
Future Work and Conclusion

ICore was initially developed to nominate documents to the appropriate step in the Physical Model. It relies heavily on nuclear concepts and terms defined in the physical model and INIS ontology.

Individual companies, people, etc. often may be entities of interest to an individual analyst, or constitute a network associated with a specific reactor, site of interest, or any entity. Future versions of ICore are designed to additionally monitor and evaluate documents being processed to identify any of these entities of interest and use this information to enhance matching to the physical model.

The source for the entities and networks may derive from ICore extracted entities that have been persisted in the IAEA Data Environment for Analysis (IDEA) prototype, or they may come from specific Palantir investigations.

Enhancing ICore to successfully perform this function will require the creation and integration of a service for resolving and disambiguating extracted individuals and companies.

ICore is scheduled to be initially deployed in an IAEA environment during the fourth quarter of 2014 with additional services (IGOR, IDIG, and IDEA) and enhancements being continuously developed and integrated throughout 2015.
Appendix 1: Nuclear Fuel Cycle and Nuclear Material Flow Diagram
i. From the IAEA Mission statement found at http://www.iaea.org/About/mission.html

ii. 1 Zetabyte = $10^{21}$ bytes or one billion TB


iv. Total number of Websites (Internet Live Stats) http://www.internetlivestats.com/total-number-of-websites/

v. The Number of Scholarly Documents on the Public Web by Madian Khabsa, C. Lee Giles. Published in PLOS, 2014

vi. For further information on the web ontology language OWL, refer to the web ontology language overview at http://www.w3.org/TR/owl-features/


viii. It is possible that two ontologies under ICore management will have logically inconsistent classes. In this circumstance, both classes will be identified and notification of the inconsistency provided is given.


x. For further reference, see Machine Learning: A Probabilistic Perspective by Kevin Murphy, The MIT Press

xi. For further information on Latent Semantic Analysis, refer to the Handbook of Latent Semantic Analysis, by Thomas Landauer and Danielle McNamara, Psychology Press; Reprint edition (June 11, 2014)


xiii. For further information of Singular Value Decomposition, refer to the text Understanding Complex Datasets: Data Mining with Matrix Decompositions, by David Skillicorn. Chapman and Hall/CRC (May 17, 2007)

xiv. Cosine similarity provides a measure of the distance between two vectors. For further information refer to any text on linear algebra.

xv. For further information on the Apache Tika project please refer to the Apache Tika website at http://tika.apache.org/
xvi. For further information on the Apache Nutch project please refer to the Apache Nutch website at http://nutch.apache.org/

xvii. MongoDB a leading NoSQL data store. Further information can be found on the MongoDB website at http://www.mongodb.org/

xviii. SOLr is an Apache project. Further information may be found on the Apache SOLr website at http://lucene.apache.org/solr/

xix. Accumulo is an Apache big-table project. Further information may be found on the Apache Accumulo website at https://accumulo.apache.org/

xx. Neo4J is a database for storing network and graph information. Further information may be found at the Neo4J website at http://neo4j.com/

xxi. The patent application may be viewed at http://www.google.com/patents/US20140270035