SAND2018-3174C

Protecting Sensitive Textual Information Using Information Extraction and Semantic Technologies

Antonio Garcia, Jina Lee, Jonathan McClain, Craig Jorgensen, and John Lewis

Sandia National Laboratories
axgarci@sandia.gov

Abstract. Protecting sensitive information is of vital importance to the business, legal, and national security sectors. Previous approaches have focused on the use of text classification and information retrieval techniques to identify sensitive information in text. While fast and able to identify a body of text as sensitive, these techniques are often black box methods and therefore lack the ability to provide users with insights into why the text was deemed sensitive. This may be unacceptable in mission critical situations where organizations need to prohibit the inadvertent release of information such as intellectual property, attorney-client privilege information, or state secrets. This paper outlines an approach leveraging domain-specific information extraction to instantiate an ontological representation of input text that is then enriched by mapping it to a domain-specific knowledge base. SPARQL queries are then used to reason on the sensitivity of the original input text. This process is not only capable of suggesting whether a textual document is sensitive but it also provides the requisite information needed to enable quick human verification of the output.

1 Introduction

For many organizations, including those in the businesses, legal, and governments sectors, there exists information which should not and/or cannot be publicly released. This information is critical to the operations of those organizations and includes a wide variety of organizational knowledge such as trade secrets, intellectual property, export controlled information, attorney-client privilege information, nondisclosure information, and state secrets. For these organizations, it would be detrimental if competitors or adversaries where to obtain access to this data. Variations of this problem have been previously explored in the legal [10] [4] and national security [2][6] domains. Regardless of the field in which this problem emerges, the techniques used tend to be forms of text classification or information retrieval (IR). While these techniques can be quite sophisticated, they often fail to leverage vital organizational knowledge and do not offer a clear human-understandable rationale for their output.

This paper defines organizational knowledge as that knowledge describing an organization's goals, operations, strategies, and processes. The complexity of this data makes it difficult to characterize in a machine-readable manner and so it is often neglected in favor of machine learning techniques that can train on a large corpus of documents in an attempt to extract that knowledge algorithmically. This is the case with the two methods previously mentioned, text classification and IR, but under real-world conditions obtaining a corpus to train these methods can have its own challenges when:

- the sensitive information frequently changes (i.e. business strategies evolve, new secret projects start up, etc.)
- the types of protections needed are highly nuanced depending on what is being done with the information (i.e. releasing it publicly, releasing it to international partners, or simply to another arm of a large conglomerate)
- the information is rarely clean and properly labeled (i.e. documents fall into multiple classes of sensitivity or are mislabeled because, while a project was once secret, the developed product has since been released).

It is posited that in the majority of instances where sensitive information exists, organizational knowledge is well known and understood within the organization that wishes to protect it. The information tends to exist in written form, maintained by subject matter experts, or is culturally ingrained in the members of the organization. Therefore, not converting this information into a machine-readable form can be a missed opportunity.

This paper describes an approach that makes organizational knowledge central to identifying sensitive information in textual data. It does this by leveraging semantic technologies to encode organizational knowledge into an ontological model that is then leveraged to create a domain-specific information extraction (IE) system. IE is a technique in natural language processing (NLP) which extracts structured information from unstructured or semi-structured text with the general goal being to allow computation to be performed.

This enables the IE system to instantiate a new ontology, representative of the input text, that is mapped, for contextualization, to an organizational knowledge base ontology. This new ontology can then be reasoned on using SPARQL queries. The queries identify sensitive information and provide the type of sensitivity concern (trade secret, export controlled, etc.), the rationale, and the provenance. Inspiration for combining IE with semantic technologies is due in part to a suggested area of research listed in [9].

2 Domain Dataset

Due to the inherent sensitive nature of this problem's data, an analog was chosen to facilitate experimentation. In choosing the data, it had to be easily obtainable, analogous to a type of business or government project, and should describe distinct aspects that could be seen as needing protection.

Given these criteria, the dataset decided upon was a collection of textual information pulled from NASA's James Webb Space Telescope website [8]. The James Webb Space Telescope (JWST) is a space-based observatory that is a collaborative effort between the National Aeronautics and Space Administration (NASA), the European Space Agency (ESA), and the Canadian Space Agency (CSA) to develop a large near-infrared and mid-infrared telescope. This telescope is currently still under development and is set to launch in 2019.

As the JWST is a NASA project, detailed information is freely available about its capabilities, mission, and progress. This information was therefore taken and sensitivities contrived allowing for the corpus to have the appearance of a government dataset with intellectual property, export control and state secret information. The JWST website's About sections were used to construct an organizational knowledge base and the News section was used to build a separate corpus for testing.

In the construction of the organizational knowledge base ontology, some of the details found in the About sections were defined as being sensitive. The sensitive information defined included information that described the system, its components, mission, launch, orbit, capabilities, specifications, etc. For the purposes of this paper, the JWST's near-infrared and mid-infrared detectors will be focused on.

3 Ontological Modeling

The organizational knowledge base was encoded, using semantic technologies, into two ontologies, a JWST ontology and an ontology that defined sensitive information. The JWST ontology define all information about the JWST program, the observatory, launch information, mission, orbit, etc. While this information could easily be broken down into multiple ontologies that more purely encapsulate aspects of the JWST project, for experimentation purposes this was consolidated into a single ontological representation. The sensitive information ontology was however separately defined as this is believed to be reusable.

3.1 Sensitive Information Ontology

In defining what JWST information would be sensitive, it was decided that an atomic piece of information which itself can be associated with some other atomic piece of information would be called a sensitive concept. Collections of these sensitive concepts allow the identification of sensitive information within text.

As an example, assuming that the phrase *JWST is launching on an Araine 5 rocket* is sensitive information then *JWST* and *Araine 5 rocket* are the sensitive concepts. While what needs to be protected is that the JWST will be on-board and launched on a Araine 5 rocket, those two sensitive concepts alone could nonetheless imply that one is launching on the other. Therefore, simply those two sensitive concepts together can be said to be sensitive information.

It possible that this implication is not always accurate and that this could lead to several false positives. However, given the requirements for certainty around this type of data and the facts that a human's involvement would likely be required in reviewing information the false positives can be deemed acceptable.

To construct a JWST ontology with identified sensitive concepts, the first thing that needed to be defined was an ontology that would enable the identification of sensitive information within an ingested body of text. To do this, a set of classes which included Document, MainDocument and Section were created where the Document is the base class from which the other two derive. MainDocument is representative of the overall textual document and the Section is representative of paragraphs within that document.

A Sensitivity class was also defined which allows for the specification that a Document has sensitive information and which includes the type of sensitivity concern, the provenance, and the rationale. Finally, a class SensitiveInstance is defined which maintains the original text found in the text document and references a SensitiveThingPosition and a SensitiveThing class (Figure 1). The SensitiveThingPosition specifies where in the document the SensitiveThing resides while SensitiveThing provides a means to distinguish which concepts in the ontology are sensitive concepts and which are not. However, this does not preclude treating all concepts equally as they are all of type `owl:Thing`. Identifying which concepts are sensitive concepts will be useful later in making the IE system domain specific.

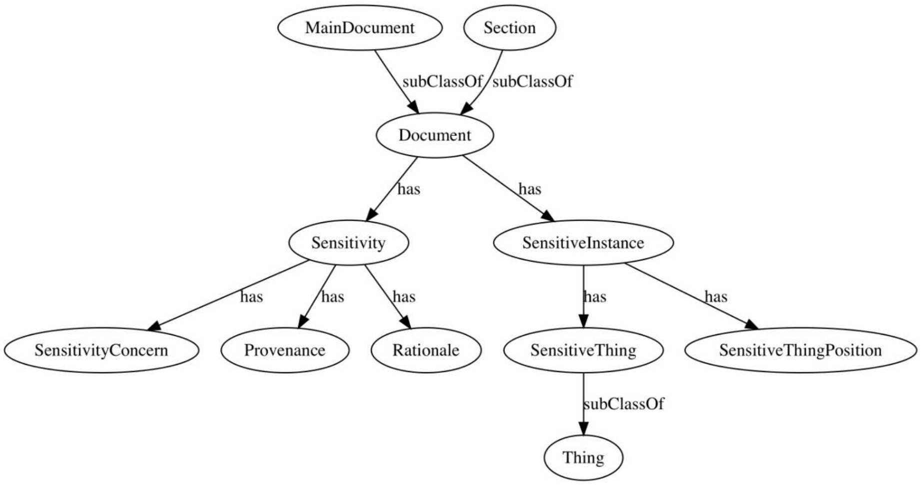

Fig. 1. Document Ontology Representation

An example of SensitiveThing usage would be if an ontology where to provide information on multiple space agencies including NASA, ESA, CSA, and the Japanese Aerospace Exploration Agency (JAXA). It can be defined that any agency having worked on the JWST needs to be identified as a sensitive concept (typed as a SensitiveThing) and those would include NASA, ESA, and

CSA. JAXA, however, is not a sensitive concept so it would not need to be a SensitiveThing (Figure 2).

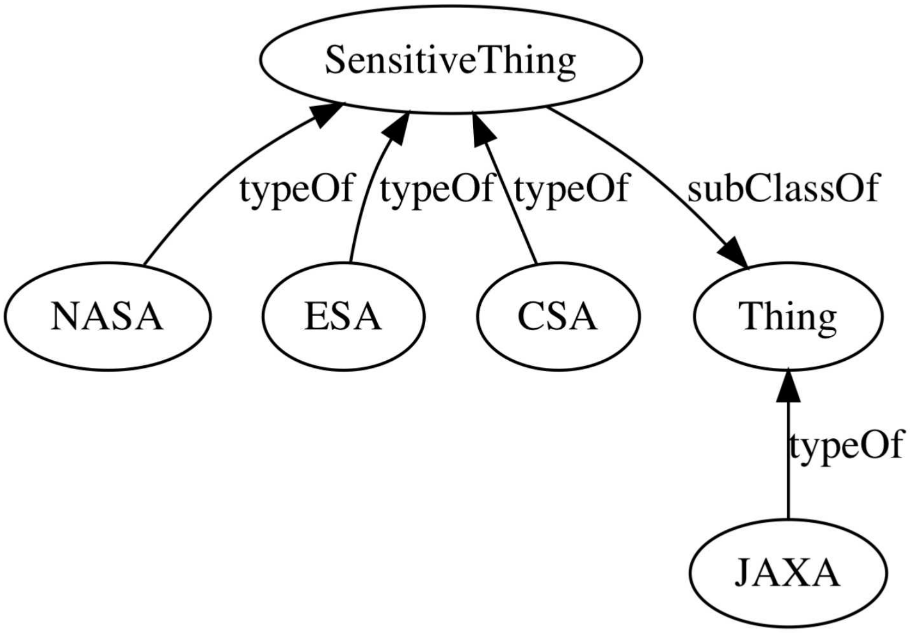

Fig. 2. Tagging Concepts as Sensitive

3.2 JWST Ontology

The JWST is a highly advanced scientific instrument with complex information related to the observatory itself, its launch, orbit, etc. Given this, the amount of information that could be encoded into an ontological model was vast, pulling from a wide range of scientific and engineering disciplines as well as administrative, governmental, and logistical knowledge. To limit what was modeled the focus was solely the protection of information deemed to be sensitive. The JWST ontology thus broadly defined the entire program and elaborated only on those areas where there existed sensitive information.

To protect sensitive concepts within the ontology, these subclassed SensitiveThing either directly or somewhere within their type hierarchy. This was important as instances of the SensitiveThing were extracted and leveraged by the IE system to make it domain-specific. The IE system looked at each instance obtained and used the SKOS [7] ontology's `skos:prefLabel` and `skos:altLabel` properties to further inform its named entity recognition system. By using the skos:prefLabel property and the skos:altLabel property multiple name variation could be input so that even variations appearing with low frequency could be identified. An example of this is the sensitive concept JWST whose preferred name was encoded to be James Webb Space Telescope but its alternative names included JWST, Webb, Next Generation Space Telescope, NGST (Figure 3). While it is possible to refer to any given sensitive concept in multiple ways, it is posited that, within an organization, there are cultural norms that limit this variation to some degree. Nonetheless, this type of information is subject to change over time and so the domain ontology is expected to require continuous curation.

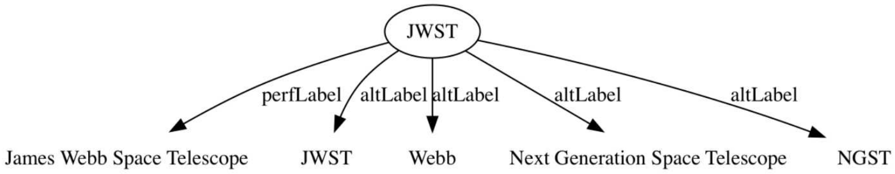

Fig. 3. JWST Labels In Ontology

Taxonomic information was another form of information that was taken advantage of by the processing pipeline. As an example of how this was leveraged, JWST has 4 infrared detectors, however 3 of those detectors are near-infrared detectors and the other is a mid-infrared detector. This information was modeled so that the individual sensors were subclasses of either MidIRDetector or NearIRDetector which themselves were both subclasses of IRDetector and that was a subclass of Detector (Figure 4). Given this example, when constructing machine-readable sensitivity guidelines, which are the guidelines that identify sensitive information using SPARQL, a single guideline could specify IRDetector and so trigger for any one of the 4 infrared detectors: MIRI, NEARCam, NEARSpec, and FGS_NIRISS.

In modeling the complexities of the JWST observatory, it was also useful to capture how components (or concepts) are associated to, or interact with, one another. Two forms of these relationships were therefore encoded - compositional and interoperability relationships. Compositional relationships were modeled to capture the structural knowledge of how a larger system is put together. This enabled inferring that a mention of a subcomponent is implying that the broader system is being referred to. As an example, mention of the interconnects layer can be said to be referring to the JWST since the JWST infrared detectors use a sandwich architecture in which the middle layer is called the interconnects layer. This example can be expanded on to define interoperability relationships which represent expressions of intercommunication, interoperation, or architectural knowledge. Along with the interconnects layer the JWST infrared detectors contains an absorber layer and a readout integrated circuit layer that

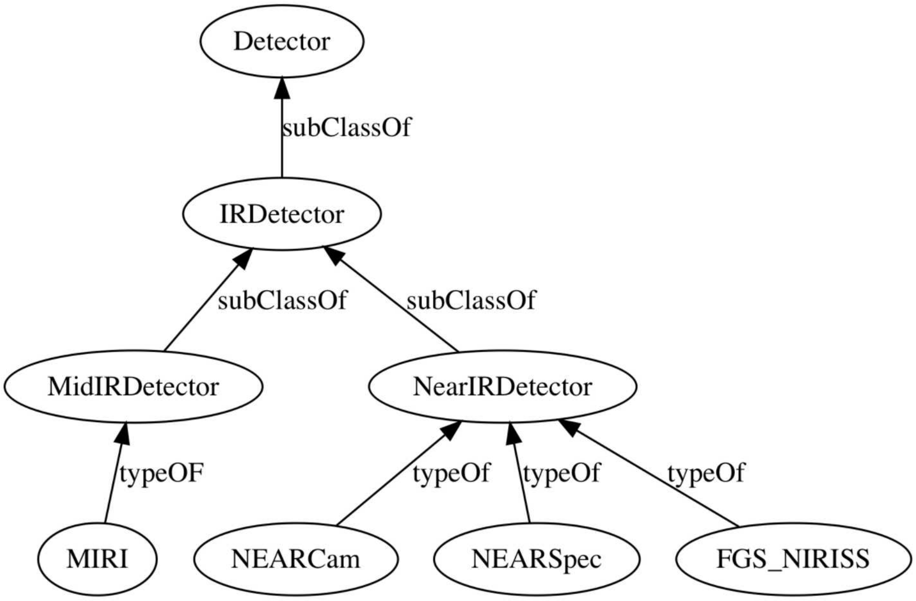

Fig. 4. Taxonomic Structure of Detectors

are on either side of the interconnects layer in the sandwich architecture (Figure 5). Knowledge of how these components interoperate, or even the fact that a subcomponent exists, potentially gives away sensitive information.

Specifications and component attribute information provided a richer contextualization of concepts being modeled. This information can be leveraged by the domain-specific IE system to enable reasoning on whether a sensitive concept is being referred to without having explicitly stated the concept. As an example, in the ontology, the JWST's primary mirror includes whether the mirror is segmented, the number of segments, the material that the mirror was made of, its size, etc. (Figure 6). Using these, the system could determine that some number of these attributes equates to a reference to the sensitive concept, primary mirror.

Lastly, human-readable descriptions were added as comments to all sensitive concepts which help contextualize identified sensitive information. For instance, in a textual document that states *the development of MIRI was completed*, MIRI may be identified as a sensitive concept and be part of some flagged sensitive information in the document. In this instance, the user may appreciate this information but, depending on their background, may not fully understand what MIRI is. If they didn't know that MIRI is one of the infrared detectors that is on the JWST, namely the mid-infrared instrument, they might question the result. Having descriptive information allows for this to be presented to the user to aid in their understanding.

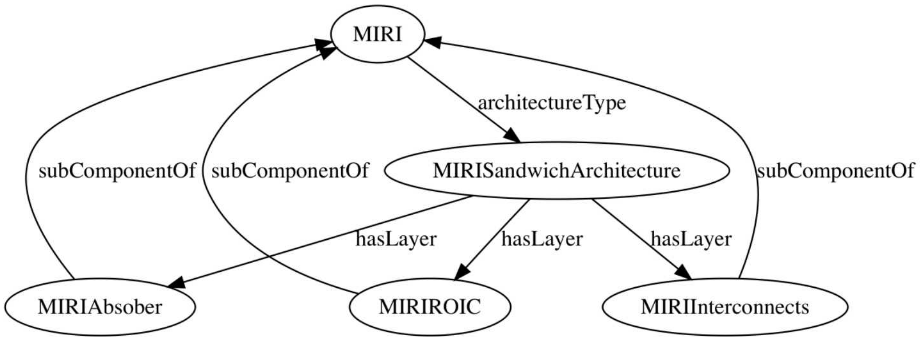

Fig. 5. Sandwich Architecture - positioning of the sandwich architecture layers has been omitted here for brevity

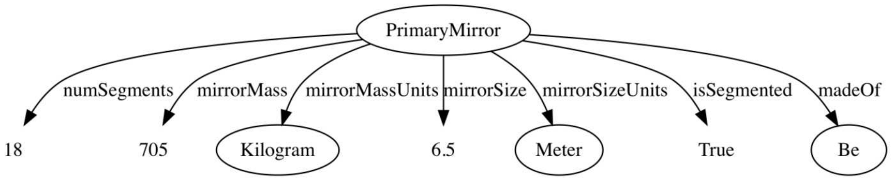

Fig. 6. Attribute and Specification Ontology

4 Domain-specific Information Extraction and Ontological Mapping

IE is a technique in natural language processing (NLP) which extracts structured information (facts) from unstructured or semi-structured textual documents with the general goal of allowing computation to be performed. Current research in this area has heavily focused on Open Information Extraction (OIE), given the advent of TextRunner [1] which extracts potential facts from the information being searched. Because it is believed that organizational sensitive information is known, the focus here is on domain-specific IE. This form of IE leverages a knowledge base that is comprised of domain knowledge to extract facts relevant to the given domain in a targeted fashion.

In general, IE outputs data in the form of triples, or n-ary propositions, which can then be mapped to resource description framework (RDF) triples to construct of a new ontology representative of the ingested text. To do this, information is pulled from sensitive concept instances in the domain knowledge base to construct NLP models. This allows for the domain-specific IE system to tag the extracted facts with URI's from the ontology and then use those to instantiate a new semantic graph representative of the ingested text. The IE processing described below used the Stanford's CoreNLP system [5] which provides a set of natural language processing tools that are highly configurable.

4.1 Ingesting Ontological Data and Named Entity Resolution

Out of the box the CoreNLP system provides named entity resolution (NER) models built from several corpora however these models are often insufficient for specialized domains because those domains often use their own names, acronyms, initialisms, jargon, etc. The CoreNLPs NER annotator recognizes named (PERSON, LOCATION, ORGANIZATION, MISC.), numerical (MONEY, NUMBER, ORDINAL, PERCENT), and temporal (DATE, TIME, DURATION, SET) entities and annotates them with the relevant NER class annotation.

Along with these entities, this NER system should also be able to recognize domain specific sensitive concepts to enable reasoning on them. As the organizational knowledge base ontology will already be used for this purpose, it can also be used to build a domain specific model for a second custom NER annotator. This custom NER annotator is intended to target domain-specific sensitive concepts and will annotate them with both the SENSTIVECONCEPT NER class and the URI of the concept in the ontology.

The sensitive concepts are extracted and put into tab-separated file where the first column contains the preferred name. The next column contains the ontological URI - the unique identifier that will enable mapping text back to the concept in the ontology. Lastly, there are *N* alternative names allowed capturing the values of all instances of `skos:altLabel`.

4.2 Coreference Resolution

Coreference resolution is the task of determining which expressions refer to the same entity in text. By performing this process it can be determined, from the two sentences below, that the ”It” referred to in the second sentence is actually the JWST.

> *JWST needs extraordinarily sensitive detectors to record the feeble light from far-away galaxies, stars, and planets. It needs large-area arrays of detectors to efficiently survey the sky.*

The CoreNLP coreference resolution annotator was configured to use the statistical system. This system is a mention-ranking model that uses a large set of features and operates by iterating through each mention identified in the document, possibly adding a coreference link between the current one and a preceding mention [3].

4.3 Information Extraction

The last step in the IE processing pipeline is to perform information extraction to extract facts from the ingested text in subject, predicate, object form. To do this, the CoreNLP OpenIE annotator is used but while this annotator is focused on open information extraction the addition of domain-specific annotations make the output triples specific to the JWST domain and capable of being mapped to the organizational knowledge base ontology.

5 Building a Model From Ingested Text

The output of the IE system can now be analyzed and a new ontological graph, representative of the ingested text, constructed. This was done by first instantiating a new MainDocument and a set of Section classes for each paragraph within the document. The annotated text is then taken and for each sensitive concept identified a SensitiveInstance is instantiated. These SensitiveInstances reference three pieces of information:

- the string representative of how sensitive concept was referenced in the text,
- a SensitiveTermPosition indicating where the term was in text, and
- the ontology’s URI of the sensitive concept

As an example, assume that the material make up of an infrared detector’s subcomponents is sensitive information and that the following are sensitive concepts: absorber layer, HgCdTe, Si:As, indium, interconnects layer, and ROIC. Upon being processed, this text below should yield the ontology in Figure 7.

> *The pixelated absorber layer (HgCdTe or Si:As) absorbs the light and converts it into voltages in individual pixels. The indium interconnects join pixels in the absorber layer to the ROIC.*

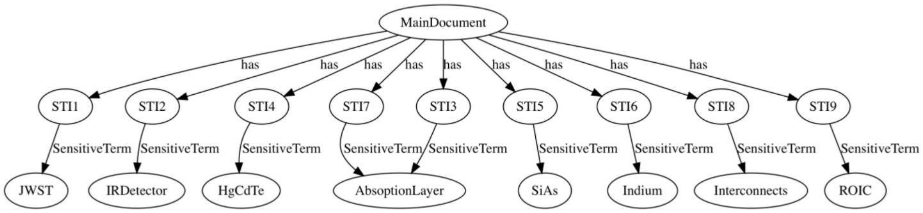

Fig. 7. Output Ontology

For readability SensitiveThingPosition and actual mentions in text are not shown.

6 Identifying Sensitive Information with SPARQL

With the ingested text having been converted to an ontological graph of sensitive concepts, this information can now be queried to determine if sensitive information is present. For the JWST, the following guidelines can be assumed:

1. $(AbsorberLayer \wedge Material) \Rightarrow HighlySensitive$
2. $(InterconnectsLayer \wedge Indium) \Rightarrow ExportControlled$
3. $(IRDetector \wedge Material) \Rightarrow HighlySensitive$

Guideline 1 indicates that mention of any of the 4 absorber layers together with any type of material is considered highly sensitive. Guideline 2 indicates that any mention of the 4 interconnects layers together with indium is export controlled. Lastly, guideline 3 indicates that any mention of an infrared detector together with any material is considered highly sensitive.

To operate on the generated ontology these guidelines are converted to SPARQL queries which can make use of the defined taxonomic, component relationships, and descriptive information implemented. Below are these guidelines in SPARQL query form.

```
SELECT ?doc
WHERE {
  ?doc cs:hasSensitiveThingInstance ?sti1 .
  ?sti1 cs:hasSensitiveThing ?st1 .
  ?st1 rdf:type/rdfs:subClassOf* jwst:Material .
  ?doc cs:hasSensitiveThingInstance ?sti2 .
  ?sti2 cs:hasSensitiveThing ?st2 .
  ?sti2 rdf:type/rdfs:subClassOf* jwst:AbsorberLayer .
}
SELECT ?doc
WHERE {
  ?doc cs:hasSensitiveThingInstance ?sti1 .
  ?sti1 cs:hasSensitiveThing jwst:In .
  ?doc cs:hasSensitiveThingInstance ?sti2 .
  ?sti2 cs:hasSensitiveThing ?st2 .
  ?sti2 rdf:type ?t .
  ?t rdfs:subClassOf* jwst:InterconnectsLayer .
}
SELECT ?doc
WHERE {
  ?doc cs:hasSensitiveTermInstance ?sti1 .
  ?sti1 cs:hasSensitiveTerm ?st1 .
  ?st1 jwst:hasSubcomponent* ?sc .
  ?sc rdf:type/rdfs:subClassOf* jwst:IRDetector .
  ?doc cs:hasSensitiveTermInstance ?sti2 .
  ?sti2 cs:hasSensitiveTerm ?st2 .
  ?st2 rdf:type/rdfs:subClassOf* jwst:Material .
}
```

Though these SPARQL queries will return a document when the criteria is met, this has yet to map which security concern is represented, the provenance, and human-readable rationale. To do this, the SPARQL queries were defined together with this information allowing a human user to be presented with all of the information for why the system believes sensitive information exists. Moreover, given that the concepts identified are in the ontology, the user can refer to them for further contextual information. Additionally, by applying these machine-readable sensitivity guidelines against a document broken down by paragraph, the system can identify if specific paragraphs are giving away sensitive information along with any sensitive information that spans multiple paragraphs.

7 Conclusion

This paper has described a method for using semantic web ontologies, IE, and SPARQL queries to aid in identifying sensitive information in textual documents. The approach can be used for the purposes of protecting an organization's trade secrets and intellectual property, to ensure data such as export controlled information is not released, to protect attorney-client privileged information, and for aiding governments in protecting their sensitive information.

Going forward, the plan is to more fully utilize facts output by the information extraction system to interpret more complex relationships in text. This would include making better use of IE predicates to reduce false positives and to attempt to capture implied sensitive concepts.

As a further goal, the plan is to look at other data modalities (image, audio, etc.) since by encoding the organizational knowledge base as an ontology, this data can be reused by other methods. An example would be if a document with imagery was submitted and the imagery potentially holds sensitive information. In this case, the processing pipeline could instantiate any sensitive concepts identified in text but also instantiate sensitive concepts that a computer vision algorithm has identified. The output of both the IE system and the computer vision algorithm would be a set of sensitive concepts in a single resultant ontology that can be reasoned on.

References

1. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence. pp. 2670–2676. IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007), http://dl.acm.org/citation.cfm?id=1625275.1625705

2. Brown, J.D., Charlebois, D.: Security classification using automated learning (scale): optimizing statistical natural language processing techniques to assign security labels to unstructured text. Tech. rep., DEFENCE RESEARCH AND DEVELOPMENT CANADA OTTAWA (ONTARIO) (2010)
3. Clark, K., Manning, C.D.: Entity-centric coreference resolution with model stacking. In: Association for Computational Linguistics (ACL) (2015)
4. Cormack, G.V., Grossman, M.R.: Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In: Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval. pp. 153–162. SIGIR '14, ACM, New York, NY, USA (2014), `http://doi.acm.org/10.1145/2600428.2609601`
5. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), `http://www.aclweb.org/anthology/P/P14/P14-5010`
6. McDonald, G.: A framework for enhanced text classification in sensitivity and reputation management. In: Proceedings of the 6th Symposium on Future Directions in Information Access. pp. 59–61. FDIA '15, BCS Learning & Development Ltd., Swindon, UK (2015), `https://doi.org/10.14236/ewic/FDIA2015.15`
7. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. Working draft, W3C (2008), `http://www.w3.org/TR/skos-reference`
8. NASA/JWST: James webb space telescope (Nov 2016), `https://jwst.nasa.gov/`
9. Piskorski, J., Yangarber, R.: Information Extraction: Past, Present and Future, pp. 23–49. Springer Berlin Heidelberg, Berlin, Heidelberg (2013), `https://doi.org/10.1007/978-3-642-28569-1_2`
10. Roitblat, H.L., Kershaw, A., Oot, P.: Document categorization in legal electronic discovery: Computer classification vs. manual review. J. Am. Soc. Inf. Sci. Technol. 61(1), 70–80 (Jan 2010), `https://doi.org/10.1002/asi.v61:1`