# Finding and Categorizing Recurring Reports to Congress

Laura Odell, Katharine Burton, and Miranda Seitz-McLeese

**The Challenge: DoD had no single source listing the recurring reports that DoD was responsible for submitting to Congress.**

**Because National Defense Authorization Acts (NDAA) and other public laws add, modify, or remove reporting requirements, manually maintaining an updated list of reports would also require significant effort and time.**

## Background

DoD is required to send multiple reports to Congress, and it is difficult to keep track of them all – when they are due, what they must contain, which office receives which report. Attempting to manually track each report would require months of sustained work. Because National Defense Authorization Acts (NDAA) and other public laws add, modify, or remove reporting requirements, manually maintaining an updated list of reports would also require significant effort and time.

The Office of the Under Secretary of Defense for Acquisition, Technology and Logistics (OUSD(AT&L)) asked IDA to identify sections of Title 10, U.S. Code that imposed recurring reporting requirements on DoD, as well as the frequency of the reporting requirements.

## Methodology

The U.S. Code is available online in XML format. We split the XML-structured version of Title 10 into sections – several thousand sections at the start – and began to identify the sections that imposed reporting requirements. In machine learning, classification is a means of determining whether an object (in this case, a section of Title 10) belongs to a certain set (or group of related objects). Classification is a common machine learning task, but most classification algorithms require a training data set before they can be applied. This task did not have a training data set. And, because the sponsor had requested IDA to minimize manual effort, manually flagging several thousand documents as “imposing requirements” or “not imposing requirements” was not practical.

Instead, we used regular expressions[1] to find a small subset of documents that impose reporting requirements. Although this subset was too small to use as a training data set for a

[1] A regular expression is a special text string that describes a search pattern (RegularExpressions.info, http://www.regular-expressions.info/. Accessed September 26, 2017).

robust classification algorithm, it was large enough for the researchers to conduct a meaningful statistical analysis. We used Bayesian techniques to identify words and phrases that were statistically more likely to indicate a reporting requirement. The researchers then wrote a simple classification algorithm based on the results of the analysis.

Once we identified the sections that imposed reporting requirements, we began extracting metadata, including report frequency, subject, and responsible office. We sorted documents by extracting terms associated with frequency and periodicity (e.g., "annual," "quarter"). We then took a sample of the remaining documents and sorted them according to those extracted terms. These phrases were added to the extraction, and the process was repeated until only a few documents remained, which had to be manually sorted.

We used Title 10's chapter headings as descriptions of the subject matter of the required reports. We manually collected the headings to create a starter training data set, and used a label propagation algorithm to group the reports under general topic areas. We then extracted the office responsible for each report using a function designed for a previous project. The function uses regular expressions and string matching to identify agencies and offices under the Secretary of Defense. The researchers created a table listing the text, citation, subject, topic area, and periodicity of the reporting requirement from each section of Title 10 and submitted it to the sponsor.

## Results

IDA found about 200 sections of Title 10 that imposed a recurring reporting requirement. The majority of these were annual reports, although biannual and biennial reports also figured prominently. Quarterly and quadrennial reports were the least frequent. Following annual reports, the most common type was event-triggered reports, or reports that required submission to Congress after a particular event occurred. Event-triggered reports are the type that DoD is most likely to lose track of, especially if the events triggering the report happen rarely.

## Impact

Before IDA's analysis, no single source listed the recurring reports that DoD was responsible for submitting to Congress. IDATA's automation capabilities saved manpower and resources: the effort was conducted by a single analyst and a few subject matter experts over a few weeks. The process used to generate the report list could also easily be adapted to update an existing list.

DoD can use the report list to improve allocation of scarce resources to the reports that need immediate attention and be prepared to tackle event-triggered reports, avoiding surprises.

## References

Glickman, M.E. and D.A. van Dyk,. "Basic Bayesian Methods," in W.T. Ambrosius, (Ed.), *Methods of Molecular Biology, Vol. 4: Topics in Biostatistics.* Totowa, NJ: Humana Press, Inc. 2007.

RegularExpressions.info. http://www.regularexpressions.info/. Accessed September 26, 2017.