KNOWLEDGE ENGINEERING TOOLS – READY TO SUPPORT RISK-INFORMED DECISION MAKING?

Nathan Siu, Margaret Tobin, Peter Appignani, Kevin Coyne

U.S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research:
MS CSB 4A07M, Washington, DC, 20555, Nathan.Siu@nrc.gov

Gary Young
U.S. Nuclear Regulatory Commission Office of Information Services:
MS OWFN 6D3, Washington, DC, 20555

Scott Raimist

ECM Universe: 13623 Bare Island Dr, Chantilly, VA, USA, 20151

The U.S. Nuclear Regulatory Commission (NRC) is performing a feasibility study on the application of advanced knowledge engineering tools and techniques to support the improved and expanded use of risk information. This work is being undertaken as part of the NRC's Long-Term Research Program. The project is pursuing a number of demonstration applications of content analytics software being explored at the NRC for a variety of situations. The project applications involve the identification and characterization of multi-unit events, the identification and characterization of common cause failure events, the characterization of current PRA results, and the characterization of severe accident design features of new reactor designs. The multi-unit application, which is ongoing, has demonstrated the usefulness of content analytics technology and some of the challenges in employing this technology. Work continues on this and the remaining applications.

I. BACKGROUND AND OBJECTIVE

As discussed by Siu et al, the U.S. Nuclear Regulatory Commission's (NRC) Office of Nuclear Regulatory Research (RES) has initiated a feasibility study to explore the application of advanced knowledge engineering tools and techniques to support Probabilistic Risk Assessment (PRA) activities.[1] The study is being performed under auspices of the NRC's Long-Term Research Program (LTRP), which is used to investigate topics expected to meet critical mission needs in 5 to 10 years.[2] Consistent with the scope of the LTRP, the project is being conducted as a scoping study aimed at determining if additional agency effort to develop actual KE tools aimed at supporting risk applications is worthwhile.

The feasibility study has two principal lines of effort. The first addresses currently available software tools used to extract information from large, unstructured information bases. In particular, the work explores the potential of "content analytics" tools, discussed in the remainder of this paper. The second line of effort addresses ongoing, non-NRC activities, notably the Open PSA initiative[a] and related projects,[3] aimed at improving current PRA modeling and documentation practices. For NRC staff, who may need to review a PRA model in the course of a risk-informed decision making process, an intriguing aspect of such efforts is that they are aimed at developing standardized representations of models which could facilitate comparisons between different models of similar systems.

This paper discusses the current status of the study team's efforts involving content analytics software.

II. CONTENT ANALYTICS

Referring to *Webster's*, the term "content analysis" is defined as the "detailed study and analysis of the manifest and latent content of various types of communication … in order to ascertain their meaning and probable effect" and the term "analytics" is defined as "the science of analysis."[4] In the information technology world, where increasing amounts of resources are being spent to make better use of large (and ever-increasing) amounts of unstructured information, "content analytics" methods and tools are being used to, among other things, help users improve their searches and enhance their "discovery" activities (i.e., activities to develop insights through exploration of the available information).

Based on responses to an NRC sources sought notice,[5] it appears that there are several companies with

[a] See www.open-psa.org for information on this initiative.

content analytics capabilities and products. This project is using an International Business Machines (IBM®) product, Content Analytics Version 2.2, as a representative of available tools.

As discussed by Zhu et al.,[6] the IBM Content Analytics tool consists of a number of major software components, including:

- "crawlers" which go through the documents in a pre-selected set (called a "corpus" by IBM) and extract document content;
- document processors which convert the unstructured text data generated by the crawlers into structured data using rules provided by text analytic "annotators" (including standard annotators to do such things as identify the document language, perform a linguistics analysis, and identify text patterns using user-supplied rules, as well as any additional custom annotators);
- an indexer which prepares an optimized index of the processed document content (called a "text analytics collection," or "collection" for short) suitable for high-speed text mining and analysis; and
- a text miner application which provides the user interface enabling an analyst to search the corpus.

The use of the Content Analytics tool requires software engineers to configure the tool (e.g., to control how a crawler uses system resources and when it should be run) and develop desired annotators, and Subject Matter Experts (SMEs) to work with the software engineers to define the search problem of interest and ensure efficient tool development, as well as the analyst user(s).

III. STUDY DESIGN

The objective of the content analytics portion of our feasibility study is to investigate the current state of content analytics technology with respect to the staff's risk-informed activities. To meet this objective, we are employing a case-study approach that applies our example tool (IBM Content Analytics Version 2.2) to a number of prototypical problems ("use cases") faced by staff. Through our evaluation of the use cases, including a comparison of the content analytics effort and results with those associated with approaches and tools currently used by staff, we expect to develop lessons supporting our overall assessment.

The specific use cases currently included in our project plan are summarized in Table I. Two of these use cases, notably Use Case 1 (investigating multi-unit events) and Use Case 2 (investigating common cause failure events), appear to be similar in nature to public demonstrations of content analytics technology (e.g., an analysis of medical device failure data collected by the U.S. Food and Drug Administration's MedWatch Program and an analysis of product defect and recall data collected by the National Highway Traffic Safety Administration). They also share similar features with activities performed by the NRC's Office of Nuclear Reactor Regulation aimed at characterizing inspection findings and how these findings are used by staff. The other two use cases (investigating current PRA results and the severe accident features of new reactor designs) appear to be further afield and may not be well matched for the tool. In all cases, we are interested in determining whether the content analytics tool answers the principal question of interest, how much effort is required to develop the answer, and what are some of the side benefits of using the tool.

TABLE I. Project Use Cases

No.	Description	Notes
1	Search for multi-unit events	Supports characterization of past events involving multiple units at a site. This characterization could support the ongoing development of a site-wide PRA model.
2	Search for common cause failure (CCF) events	Identify and characterize past CCF events. The results of this activity could support the conduct of an expert elicitation for the likelihood of events potentially relevant to the analysis of Interfacing Systems Loss of Coolant Accidents (ISLOCAs).
3	Characterization of current licensee PRA results	Support decision maker understanding of current risk levels and contributors. This activity addresses a common question raised by managers and external stakeholders.
4	Characterization of severe accident design features of new reactor designs	Support staff and decision maker understanding of key features for a given design and of differences between designs.

TABLE II. Project Corpus – Current Contents

Description	Notes
Publicly available documents from NRC's Agencywide Document Access and Management (ADAMS) Main Library	Includes NRC staff (NUREG) and contractor (NUREG/CR) reports, staff papers to the Commission (SECY papers) and Commission Staff Requirements Memoranda (SRMs), License Amendment Requests, New Reactor Design Control Documents.
Final Safety Analysis Reports	Provide terminology and design-related information useful for event analysis
Standardized Plant Analysis Risk (SPAR) model documentation	Provides design-related information useful for event analysis (e.g., the size of the system involved), PRA results that can be compared with licensee/applicant results
Immediate Notifications	Documents notifying the NRC of events submitted per 10 CFR 50.72
Licensee Event Reports (LERs)	Documents notifying the NRC of events submitted per 10 CFR 50.73
Inspection reports	Staff reports from the NRC's Reactor Oversight Process (1999-present)
Individual Plant Examinations	Licensee submittals in response to Generic Letter 88-20
Individual Plant Examination of External Events	Licensee submittals in response to Generic Letter 88-20, Supplement 4
Advisory Committee of Reactor Safeguards (ACRS) letter reports	1985-present
ACRS Meeting Transcripts	1999-present (subcommittee as well as full committee)

All of the use cases use the same corpus, summarized in Table II. Note that some of the documents in Table II are not publicly available.

IV. USE CASE 1 – MULTI-UNIT EVENTS

As argued by Fleming in 2005 (Ref. 7) and illustrated by the March, 2011 reactor accidents at the Fukushima Dai-ichi nuclear power plant, events involving multiple reactor units at a single site can be important contributors to site risk. There are numerous technical challenges in assessing these contributions.[7]

NRC/RES is currently engaged in a full-scope, Level 3 PRA study intended to address all relevant site radiological sources (including the spent fuel pool and dry cask storage), internal and external initiating event hazards, and modes of operation for a 2-unit, Westinghouse four-loop pressurized water reactor station with a large, dry containment.[9,10] The technical approach for addressing multi-unit (and, more generally, multi-source) events is described in broad terms in the project's Technical Analysis Approach Plan.[11] To inform the modeling of such events, it is of interest to review past operational events to provide a sense of the likelihood and impact of these events, and of their salient features.

IV.A. Objective and Scope

The purpose of this use case is to search for and characterize past U.S. operational events involving multiple reactors. The focus of the search is on events involving a transient at one or more units at a single site. The search is not aimed at identifying degraded conditions that could affect the response of multiple units during an accident. The event characteristics currently addressed are:

- Event Date
- Facility Name
- Event Extent
- Event Cause

IV.B. Prior Work

One purpose of this use case is to compare the required effort and results of using the Content Analytics tool with those of a previous manual search. This latter search was performed in support of the previously mentioned NRC Level 3 PRA study.

The manual search used the Licensee Event Report (LER) database maintained for the NRC by Idaho National Laboratories (INL). LERs are reports submitted to the NRC from the plants in accordance with the requirements of 10 CFR 50.73. Examples of reportable events are losses of safety related systems or reactor trips. The LERs document the root cause, mitigating actions, and corrective actions that the plant has taken. The search included LERs reported over the period 2003-2013.

The manual search took under four weeks to complete. The general process used the following steps.

- Search the INL LER database to generate a spreadsheet of events for all sites that have two or more units and for which the word "trip" appeared in the report, in an effort to exclude degraded conditions.
- Sort search results by date and select all of the events that have either:

- Both unit docket numbers on the LER, OR
- Two separate LERs that were dated within 24 hours of each other at one site.

Once the list of events was compiled and sorted, the each LER was reviewed to determine if it represented a genuine multi-unit event, or if it did not (i.e., if it was a "false positive"). An example of a multi-unit event that this search found was the multi-unit loss-of-offsite-power at Nine Mile Point Station and at Indian Point during the 2003 East Coast blackout. Examples of false positives include occasions where the LER listed both docket numbers (i.e., the LER was "dual docketed") but the second unit was not called out in the text of the report, or only called out to say that there was no effect on the unit, or where two events from a single unit occurred within the search-specified 24 hour timeframe.

From the initial results, there were approximately 50 initial hits that met all of the requirements listed above. Upon further review, it was determined that about 30-40 percent of the results were false positives, leaving about 30-35 actual multi-unit events.

In general, the manual search identified several loss of offsite power (LOOP) affecting both units, events where one unit was in shutdown with equipment out of service when the other unit had an event, and a number of events that seem to be plant specific (e.g., a specific plant layout caused the event to affect the other unit). The LOOP events were the most common single event, but many of them came from a single blackout (the East Coast blackout of 2003).

IV.C. Content Analytics Analysis Method and Tool

The principals involved on Use Case 1 are a PRA engineer who had performed the manual search described in the preceding section and a pair of software engineers expert in the development of Content Analytics applications. At the beginning of the use case work, the software engineers were provided an initial specification of the use case. Based on this specification, they developed annotators and a text miner application to address aspects of the specification that were addressable within available project resources. As of the writing of this paper, they are interacting with the PRA engineer to improve the annotators and text miner. In general, the PRA engineer's focus is on ensuring the clarity of the search problem and the practical usefulness of the Content Analytics results, whereas the software engineers' focus is on developing the tool enabling the PRA engineer to perform the search.

It's worth noting that the original problem specification for this use case included an objective to determine the following characteristics for identified multi-unit events.

- Plant(s)/licensee(s)
- Event time/date(s)
- Proximate cause
- Extent (including system/structure/component – SSC – degradations as well as complete failures)
- Sequence of events
- Safety significance (e.g., conditional core damage probability – CCDP)
- Corrective action(s)
- Contextual information (e.g., if triggered by a flood, was the flood a beyond design basis event?)

Comparing this list with the list provided in Section IV.A., it can be seen that a number of potentially interesting characteristics are not being addressed in this project. (Of course, the descriptions of identified events can still be manually reviewed for desired information.) The tool development process involved discussions between the software engineers and the PRA engineer to help specify exactly what information each characteristic was intended to convey.

An example of a characteristic that required some back-and-forth discussion was the event "proximate cause." Not surprisingly, because the identification of a proximate cause requires a certain amount of model-based reasoning (to specify what was the "last event" in a causal chain prior to the actual multi-unit event), a significant amount of work would have been required to apply the general-purpose natural language tools provided with the Content Analytics software. Ultimately, it was decided that developing an automated proximate cause identification process would be too resource intensive for the purposes of our technology evaluation study. Therefore, the proximate cause characteristic in the original specification was replaced with a more general cause characteristic and the tool supports the search for any cause in the causal chain, including the proximate cause and root cause.

One of the event characteristics that was surprisingly difficult to implement was the event date. The LERs (currently in the form of pdf files) all use a form with a graphically delineated (boxed) field for the date which is immediately recognizable to a human reader. However, when the LERs are scanned and converted to electronic documents using optical character recognition (OCR), the particular document processors used by the Content Analytics software are unable to easily distinguish the string of numbers and separating symbols representing a date from other strings of numbers and symbols. Furthermore, errors in the scanning and OCR (e.g., due to aging of or even errors in the hard copy) raise additional difficulties. Given the importance of event date in the search scheme, a large amount of time and programming effort went into developing a custom routine that was able to identify the LER dates.

From a PRA perspective, most of the work performed by the project's software engineers is "behind the scenes." For example, the PRA analyst and other end users generally do not construct or perform a detailed review of the annotators produced by the software engineers. Rather, they use a customized text mining application, also produced by the software engineers, which provides a number of tools supporting user searches and discovery. The principal tools are "facets," different windows on the corpus data, and their associated searches. (So, for example, clicking on a facet tab for Multi-Unit Events will bring up the results of a search through the corpus based on the annotator rules associated with Multi-Unit Events.) Other tools can filter search results and support the development of statistics (e.g., matching document counts, frequencies of and trends in search phrase occurrences, and correlations of pairs of search phrase occurrences) and the visual identification of database relationships between facets.

Work is underway to improve the current version of the search tool. This work includes the provision of a more complete list of SSC names to aid searching, and the development of refined rules to eliminate common false positives. For example, initial searches using the term "trip of both reactors," found events involving the "trip of both reactor protection trains," and these included events involving the trip of both trains in a single unit. Other situations similar to this one still exist in the model, but fine tuning these results is an iterative process, and false positives have to be removed as they are discovered.

IV.D. Current Results

From an initial corpus of approximately 115,000 documents, our Content Analytics tool has identified approximately 700 potentially relevant documents. A quick review of several (approximately 100) documents, using the tool's ability to display a few key sentences of the relevant portions of a document without actually opening the document, shows that about 60 percent appear to be false positives. Trends found in these false positives will be used to refine the tool to reduce the rate of false positives and improve the efficiency of the search. More information on the types of false positives found is provided in Section IV.E below.

When using the content analytics tool to address the same time period covered by the manual search discussed in Section IV.B, the content analytics search resulted in about 80 hits. Thus, there were more hits from the content analytics search, but also a higher incidence of false positives. However, the content analytics tool has the benefit of not being limited to the LER database, and can search from other sources, such as inspection reports. This is a major benefit, because the manual search quickly becomes too large to reasonably complete.

The Content Analytics tool found several LERs that were useful for the Level 3 PRA Project, but that were not found in the manual search due to being outside the range of dates covered by the manual search. The ability to readily search a broader range of documents is a major benefit of the tool. However, there also are a few multi-unit events identified by the manual search that the tool did not identify. An example of such an event occurred August 22, 2013 at Arkansas Nuclear One where, during lifting and removal of the main generator stator, the lift assembly collapsed. The stator fell onto the turbine deck and caused structural damage to both units' Turbine Buildings and non-vital electrical systems. This "miss" of the Content Analytics search may be due to the inherent difficulty in crafting a completely effective automatic search. Our project has shown that with a moderate resource investment, automatic search tools capable of finding matches for multi-unit events (which are relatively rare occurrences) in large, unstructured databases can be developed. However, more effort is needed to develop the complex search rules that may be needed to catch unique occurrences that don't conform to the general trends exhibited by most multi-unit events.

IV.E. Additional Notes

Case 1 has required approximately two to three weeks of effort (staff time) from the PRA engineer, and a somewhat larger level of effort from the software engineers. Much of the PRA engineer's time was involved in becoming familiar with the tool and its capabilities, and engaging in the iterative effort of getting the tool into its current form. The tool, while simple to use, had a fairly steep learning curve which requires a relatively large amount of time to understand the results and methods of creating those results. Without an understanding of the tool, it can be difficult to gather meaningful insights from the search results.

For example, when the tool is first opened, it opens on the results tab, which, since no searching has been done in the session yet, just displays all of the documents in the corpus. It is not immediately obvious to the new user that they need to click on the "Facets" tab to start a search. Once in the Facets tab, the tool displays a number of facets and options, and it is not always clear what needs to be done to start a search. For an example of what the Facets tab looks like, see Figure 1. In order to search on the multi-unit facet, all of the check boxes next to the words must be selected, and then the "And" button must be used (see inset in Figure 1).

At a more fundamental level, the PRA engineer was naturally curious as to how the Content Analytics tool framed the search problem and how it developed the solution to this problem. (For example, one question regards the literal rules built into the annotators.) Although it isn't yet clear that such knowledge is essential

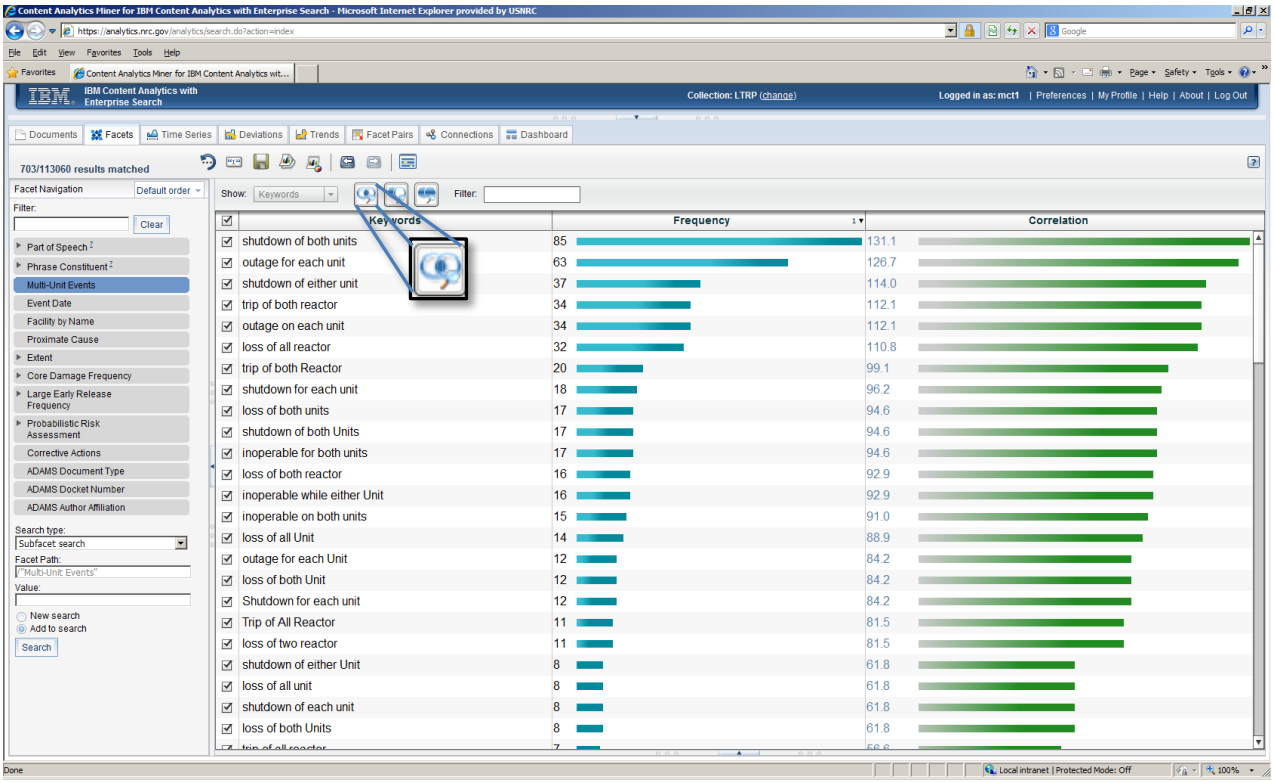

Fig. 1. Facet view of Content Analytics tool

for the SME analyst, it appears to us that the optimal level of SME engagement in the tool development process (e.g., whether the SME should be deeply involved in the development of annotators) appears to be a topic worthy of further discussion.

One of the benefits of the Content Analytics tool is its ability to leverage past applications to current needs. While the initial creation of synonym lists, facets, and annotators is fairly resource intensive, once a list is created it can be modified to fit future needs easily. For example, a complete list of systems from the various plants, once completed, could be used in any future facet. We note that our Use Case 1 benefited from a previous NRC project using Content Analytics to characterize inspection findings, and that our Use Case 2 (addressing CCF events) should benefit from the work done on Use Case 1.

A challenge of this project involved communication within the team. The software engineers fully understood how the Content Analytics tool works, but did not have the level of knowledge about LERs and how nuclear power plants function that the PRA team members have, and the PRA team members did not (as discussed above) understand exactly how the tool worked. Challenges arose from different interpretations of words. For example, in reviewing the initial search results, the team realized that there was a miscommunication in the definition of the term "block valve." The software engineers interpreted this to mean "a valve that is blocked" rather than a valve capable of cutting off flow when necessary.

Finally we note that as in the case of many analytical exercises (including PRA), the act of developing a multi-unit search tool has provided learning benefits. For example, attempting to develop the "Proximate Cause" facet gave members of the team an opportunity to think critically about how to determine what the proximate cause really is, and how to determine which cause is proximate to the event. It also demonstrated that the root cause is much easier to determine from the text of a report, but the proximate cause can be just as informative and interesting.

V. CONCLUDING REMARKS

This paper provides: a) an overview of an NRC scoping study exploring the application of advanced knowledge engineering tools and techniques to support PRA activities, and b) a status of one of the study's tasks addressing the use of content analytics. This task, which is aimed at identifying and characterizing multi-unit events, has demonstrated the usefulness of content analytics in searching through a very large number of documents to identify events not identified in a previous, more limited manual search. The task has also illustrated some of the challenges in using a content analytics approach, including the effort required to address event

characteristics of interest (e.g., proximate causes) and to develop rules that will capture all events of interest.

To date, we have focused our efforts on database searches, and have not explored content analytics tools (including trending tools, correlation analyses, and concept relationship charts – "bubble charts" – that might provide further insights regarding the available data and aid database discovery efforts. In the future, in addition to developing improved rules to reduce the number of false positives, we will spend some effort on investigating the potential benefits of these tools. We will also perform a comparison of results with those developed by Schroer and Modarres in their earlier analysis of multi-unit events.[12]

ACKNOWLEDGMENTS

The authors gratefully acknowledge the software engineering support provided by K. Bojja (ECM Universe) and the helpful comments of K. G. Golshan, D. Halvorsen, T. Nakanishi, and B. Diehl (NRC) during project discussions.

REFERENCES

1. N. SIU, P. APPIGNANI, and K. COYNE, "Knowledge engineering tools – an opportunity for risk-Informed decision making?" *Proc. Intl. Topical Mtg. Probabilistic Safety Assessment and Analysis (PSA 2013)*, Columbia, SC, September 22-26, 2013, American Nuclear Society, La Grange Park, IL (2013).
2. U.S. NUCLEAR REGULATORY COMMISSION, "Research Activities FY 2012-FY-2014," *NUREG-1925 Rev. 2*, Washington, DC (2013). (Available from the NRC Agencywide Document Access and Management System – ADAMS – Accession Number ML13242A030)
3. M. HIBTI, T. FRIEDLHUBER, and A. RAUZY, "Automated generation of event trees from event sequence diagrams and optimisation issues," *Proc. PSAM Topical Conf. in Tokyo in Light of the Fukushima Dai-ichi Accident*, Tokyo, Japan, April 14-18, 2013.
4. *Webster's Third New International Dictionary, Unabridged*, P. B. GOVE, Ed., G. & C. Merriam Co., Springfield, MA (1969).
5. U.S. NUCLEAR REGULATORY COMMISSION, "Advanced Knowledge Engineering Tools to Support Risk-Informed Decision Making," Solicitation RES-13, July 1, 2013. (Available from www.fbo.gov.)
6. W.-D. ZHU, et al., *IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content, Second Edition*, International Business Machines Corporation, Armonk, NY (2011).
7. K. N. FLEMING, "On the issue of integrated risk – a PRA practitioners perspective," *Proc. Intl. Topical Mtg. Probabilistic Safety Analysis (PSA '05)*, San Francisco, CA, September 11-15, 2005, American Nuclear Society, La Grange Park, IL (2005).
8. N. SIU, D. MARKSBERRY, S. COOPER, K. COYNE, AND M. STUTZKE, "PSA Technology challenges revealed by the Great East Japan Earthquake," *Proc. PSAM Topical Conf. in Tokyo in Light of the Fukushima Dai-ichi Accident*, Tokyo, Japan, April 14-18, 2013. (ADAMS ML13038A203)
9. U.S. NUCLEAR REGULATORY COMMISSION, "Options for Proceeding with Future Level 3 Probabilistic Risk Assessment (PRA) Activities," *SECY-11-0089*, Washington, DC (2011). (ADAMS ML11090A039)
10. A. KURITZKY, N. SIU, K. COYNE, D. HUDSON, and M. STUTZKE, "L3PRA: Updating NRC's Level 3 PRA insights and capabilities," *Proc. IAEA Tech. Mtg. Level 3 Probabilistic Safety Assessment*, Vienna, Austria, July 2-6, 2012, International Atomic Energy Agency, Vienna, Austria (2013). (ADAMS ML12173A092)
11. U.S. NUCLEAR REGULATORY COMMISSION, "Technical Analysis Approach Plan for Level 3 PRA Project, Rev 0b" Washington, DC (2013). (ADAMS ML13296A064)
12. S. SCHROER and M. MODARRES, "An event classification schema for evaluating site risk in a multi-unit nuclear power plant probabilistic risk assessment," *Rel. Engr. Sys. Safety*, **117**, 40-51 (2013).