NATIONAL ACADEMIES Sciences Engineering Medicine

2019: Safer Offshore Energy Systems Grants 4 Final Report

GULF RESEARCH PROGRAM

Project Title: Aggregating Essential Exposure Data to Enable Meaningful Analysis of Safety Incident Rates Around the World
Award Amount: $739,992
Awardee: American Bureau of Shipping
Award Start Date: 12/01/2019
Award End Date: 05/31/2021
NAS Grant ID: A-19786

Project Director: Xiaozhi Wang
Affiliation: American Bureau of Shipping

Project Key Personnel:
[List of key personal]

- Matthew Mowrer, ABS Group
- Stein Haugen, Safetec
- Joseph Myers, ABS Group
- Benjamin Roberts, ABS Group
- Anthony Barrett, ABS Group
- David Stalfort, ABS Group
- Terje Dammen, Safetec

I. PROJECT SUMMARY (from proposal)

For many years, government agencies, industry groups, and companies have collected offshore incident data to help understand and improve safety. These datasets were collected by various sponsors over diverse data periods with different spheres of interest to inform numerous decisions. Today, these datasets largely exist in silos which complicate comparison and consolidation. Aggregating these data requires that they first be recast to use common goals, measures, units, and terminology. Furthermore, all safety incident data must be understood within the context of risk exposure so that incident frequencies can be converted to incident rates for valid comparison across data sources, regions, operations, etc.

The research team includes the necessary skills and experience in safety risk analysis, offshore domain expertise, data science, and actuarial science to support this effort. The research team has performed work in many of the data silos and have long recognized the need for a comprehensive global dataset. The project plan includes literature research and stakeholder outreach tasks to develop an in-depth understanding of available datasets. Recent advances in data science tools provide novel methods that can help make this vision a reality vision. The research team sees several potential applications, such as using categorization machine learning to group similar structured data fields across different datasets or using natural language processing to extract key tags out of unstructured narratives. The potential benefits are many, including: (1) efficient aggregation, (2) tapping into previously inaccessible data, (3)

joining disparate datasets.

II. PROJECT SUMMARY (from final report)

For decades, government agencies, industry groups, and companies have collected offshore incident data to help understand and improve safety. However, multiple organizations are collecting this data over diverse data periods with different ranges of interest to inform different decisions. This research examines the feasibility of generating more precise incident rates to improve risk-based decisions by leveraging available incident and exposure data from data sources spanning multiple countries.
Offshore Oil and Gas (O&G) production is a critical contributor to the global supply of oil and gas over the last 20 years, with production volumes remaining relatively stable over that period. During this same period, however, there has been a fundamental shift in Gulf of Mexico (GoM) oil production. As new production in the GoM continues to move to deeper water, shallow-water platforms are reaching the end of their life causing the total number of platforms in the GoM to decrease. Nowhere is that more evident than in the U.S. GoM, which has seen a dramatic shift in the nature of operations. These changes raise the issue of how safety performance and the safety risk profile will be affected.
To address these issues and effectively support risk-informed decision making, safety incidents must be understood through a wide variety of facility and incident attributes. However, it is not enough to understand the nature and frequency of past incidents, as these may simply be a reflection of the number of operations. Rather, decision-makers must understand the rate of these incidents as a function of exposure variables to better understand trends and predict future incidents to manage relevant risks.
The focus of the research was to:

- Understand the viability of aggregating incident and exposure data from relevant worldwide offshore incident & safety data into a comprehensive database to enable management of safety and environmental risks.
- Develop recommendations for a standardized incident taxonomy and an ideal incident and exposure data model and the types of decisions that could be informed by the information in order to manage risks.
- Identification of data sharing opportunities and challenges that must be addressed.
- Develop recommendations for viable data science technologies that could be employed to help facilitate the development of a comprehensive database.

III. PROJECT RESULTS

Accomplishments

An initial literature review provided a foundation for understanding the existing approaches to risk assessment and management frameworks, specific risk assessment and techniques and data collection initiatives unique to the offshore O&G environment, and relevant trends in data science technology.
To evaluate the feasibility of using incident and exposure data to better quantify and manage safety risk in offshore O&G operations, our research team originally proposed research in five areas; Identify incident/exposure data sources, research and catalog selected incident datasets, document current

state, research data science tools for integrating and normalizing datasets, and develop recommendations for follow-on pilot projects. After initial data gathering, and informed by the literature review, we revised our research approach into four steps, which included:

- Developing a data model
- Gathering and cataloging data
- transforming data as needed
- Populating a data model

Developing the data model established the data management framework for the rest of our research. Our objective in this task was to compose an ideal model that maintains the strengths of prior data collection efforts, leverages modern data architectures, and enables more granular risk management by industry and regulators. We developed a data model shaped by the decision landscape and risk assessment techniques.

To populate the data model and understand data gaps, the research team first needed to gather and catalog data. We began our data search with an examination of countries known to have somewhat similar safety cultures or geographies to the U.S. We initially look at the countries who are members of the International Regulators' Forum (IRF), which include Australia, Brazil, Canada, Denmark, Ireland, Mexico, Netherlands, New Zealand, Norway, United Kingdom, and the United States. Our team began by searching for organizations that collect offshore incident and exposure data in those countries, including government agencies, industry groups, universities, and private companies. We reached out to internal and external offshore experts from around the world that have extensive working experience with the incident and exposure datasets to ensure a comprehensive identification of all relevant data sources. There were several challenges we faced when aggregating this data. Firstly, each country reports oil and gas production in different units. These were easy to convert, but we needed to take care regarding orders of magnitude when converting. Secondly, much of this data was in PDF or Excel-based reports. We were able to use tools to convert PDF files, and the cleanup of Excel files was done manually. Finally, the biggest challenges came with compiling detailed incident data.

Data science technology is available to help transform inaccessible data, match and duplicate data, and augment and normalize data to improve data analysis and ultimately risk-based decision making. Our objective was to research viable data science technologies that could be employed to extract, clean, and standardize each dataset to achieve a normalized database. The focus of our data science research was in three areas: inaccessible data, data matching and deduplication, and augmentation and normalization of data. We applied a variety of analytic and natural language processing tools to transform data for use in the data model.

In our research proposal, we suggested that the final task be the development of recommendations for a pilot implementation based on our findings. To support our recommendations, we decided to construct a proof-of-concept prototype. The proof of concept demonstrates the potential value of the data model, data gathering, and data transformation steps when implemented as an operational solution. We constructed the proof of concept using Microsoft SQL Server along with Microsoft Power BI for analytics and visualization. Our first results focused on visual representations of the connection between exposure units from different data sources. We then created a relationship to relate exposure units across assets, such as production by facility complex, rather than well. Finally, by incorporating incident frequencies, we developed an array of metrics, charts, and graphs to provide insight into potential offshore safety risk

drivers. We connected incidents to specific offshore well activities. Although the data in this proof-of-concept have been reduced too far to generate true results, the interface immediately highlighted interesting trends and findings. It is important to realize that these are only notional conclusions; the data are too incomplete for any true risk analysis results. However, despite the limited data, the data model and associated analytics proved insightful.

The overarching objective of this research is to develop an understanding of the feasibility of building a comprehensive database that integrates exposure and incident data to inform a more insightful analysis of incident rates internationally. The scope of our research was intentionally open-ended, and our research roamed accordingly. But always with the broader vision for improved data gathering and analysis capabilities. Through this process, we found several types of gaps in the available data which would prevent them from being integrated into the ideal comprehensive database. Overall, the U.S. and Norway provided the highest value data for avoiding these pitfalls. Many of the data science methods that we tested did not turn out to be as relevant as we might have expected. We also successfully used active machine learning to quickly train a model for identifying incident types. This is a critical capability for integrating data from multiple sources with different source data taxonomies. In our estimation, the most compelling result of this research was the development of the target data model. This model, summarized in Section 5.1, was built around the types of available data, the array of relevant performance metrics that we derived during this project, and the emergence of new technology which enables more flexible data management. Most incident data management systems must grapple with how to document complex incidents with multiple causal factors and multiple event types (e.g., a crane failure that results in a dropped load and explosion). These types of events require highly precise taxonomies to ensure consistent documentation and reporting.

Implications

The overarching objective of this research is to develop an understanding of the feasibility of building a comprehensive database that integrates exposure and incident data to inform a more insightful analysis of incident rates internationally. The scope of our research was intentionally open-ended, and our research roamed accordingly. But always with the broader vision for improved data gathering and analysis capabilities. Through this process, we found several types of gaps in the available data which would prevent them from being integrated into the ideal comprehensive database, such as:

- Data do not exist
- Data are not accessible
- Data are not representative of the historical offshore oil and gas experience
- Data do not connect/integrate with relevant other data
- Data are incompatible or insufficient for reliable machine learning-based enhancements

Overall, the U.S. and Norway provided the highest value data for avoiding these pitfalls. Although we tested certain technology, such as PDF scraping, with the U.S. data, some of these data in hard-to-access formats are likely also available in digital format upon request from the data owners. Key incident data from Norway was provided by Norway's PSA after a simple request and non-disclosure agreement. These two countries also had the most comprehensive publicly available data which we could find. In addition, these data are heavily populated with unique IDs which can be cross-referenced across multiple datasets, improving the potential completeness of the integration. We also expect that the data will be highly applicable beyond their intended purposes. For example, data describing wellbores, the date of their drilling, and the rigs involved provide information about drilling speed, rig utilization, and

the hours spent drilling. These can improve the precision of exposure variables associated with drilling risks.

Many of the data science methods that we tested did not turn out to be as relevant as we might have expected, with several notable exceptions. First, we spent a lot of time testing methods of automatically matching and deduplicating multiple datasets. Ideally, a comprehensive dataset such as reported incidents should not need to be deduplicated or supplemented by additional data. However, our work with BSEE and BOEM has proven that this is an important capability when attempting to maximize the completeness of an incident dataset or reintegrate parallel versions of data that have gotten out of sync. These are real-world data issues, and we believe that our tests demonstrate that the methodology was effective.

Second, we successfully used active machine learning to quickly train a model for identifying incident types. This is a critical capability for integrating data from multiple sources with different source data taxonomies.

In our estimation, the most compelling result of this research was the development of the target data model. This model was built around the types of available data, the array of relevant performance metrics that we derived during this project, and the emergence of new technology which enables more flexible data management. Most incident data management systems must grapple with how to document complex incidents with multiple causal factors and multiple event types (e.g., a crane failure that results in a dropped load and explosion). These types of events require highly precise taxonomies to ensure consistent documentation and reporting. Furthermore, a single incident type label almost always fails to capture the nuance of the event. In the proposed data model, standard data taxonomies that break down incidents into major hazard groups (i.e., fire, explosion, spill, slip/trip/fall, dropped object, collision, equipment failure, etc.) can be directly included in event sequences that account for any and all such elements of the event. Furthermore, our proposed model aims to directly relate these events and sub-events to the relevant assets and activities involved.

Education and Training

Not applicable

IV. DATA AND INFORMATION PRODUCTS

This project produced a conceptual model for ideal data collected in the offshore environment, which is housed in GRIIDC.

V. PUBLIC INTEREST AND COMMUNICATIONS

Most Exciting or Surprising Thing Learned During the Project

The most exciting conclusion from our research is that given a common decision landscape and a common risk assessment approach and comprehensive data model can be developed and then populated to help regulators and industry players analyze incidents and develop preventative or mitigative actions to improve offshore safety.

Outcomes Achieved During the Project

We produced results in each of the four area of our research.

1) Developed a Data Model. To develop the data model, we first established two categories of decision makers that our ideal data model should serve: regulators and industry players. We then constructed a data model that included the types of decision, the decision to be made, the options and the factors involved in the decisions. We then developed a risk assessment approach using a Bowtie technique which, like a safety incident, centers around a central hazardous event. On one side of the event are preventative barriers proceeding the events. On the other side of the event are mitigative barriers and potential negative outcomes of the incident. Many descriptions of reported incidents offer insights into the threats, barriers, and consequences associated with the reported hazardous event. The inputs to a Bowtie model quantify three factors:

- The frequency of threat occurrences
- The probability of all barriers failing
- The severity of potential consequences

The Bowtie model organizes these values to reduce the risk assessment to a simple equation:
Risk = Frequency * Probability * Severity
2) Data Gathering and Cataloging. In order to populate the data model and understand data gaps, the research team first needed to gather and catalog data. As the primary focus of this study was to create a model applicable first and foremost to the OCS GoM, we focused our efforts on obtaining data from countries with either similar safety cultures or geographical conditions as the U.S. These countries included the US, Norway, the United Kingdom, Mexico, and Australia. We also gathered data from the International Regulators Forum (IRF), which was formed in 1993 to promote safety practices in the global offshore arena. There were several challenges we faced when aggregating this data. Firstly, each country reports oil and gas production in different units. These were easy to convert, but we needed to take care regarding orders of magnitude when converting. Secondly, much of this data was in PDF or Excel based reports. We were able to use Adobe Acrobat to convert PDF files, and cleanup of Excel files was done manually. Finally, the biggest challenges came with compiling detailed incident data.
3) Data Transformation. We found it necessary to employ data science technology to help transform inaccessible data, match, and duplicate data and augment and normalize data to improve data analysis and ultimately risk-based decision making. Our objective was to research viable data science technologies that could be employed to extract, clean, and standardize each dataset to achieve a normalized database. The focus of our data science research was in three areas: Inaccessible data, data matching and deduplication, augmentation and normalization of data. Due to the relatively small number of documents that need parsing for inaccessible data transformation, we were able to rely on manual download. The OCS reports consist of only seven PDF files publicly available for download from the web. The OCS reports consist of seven PDF files that are publicly available for download from BSEE's website. We found several areas in the data that benefitted from record matching. For example, we were able to link production and incident data to specific platforms using a combination of text string matching and geospatial analysis. We were also able to identify duplicate records for removal (i.e., deduplication). In the coming sections, we will discuss our application of data matching to identify potential duplicates in the USCG NRC spill incident data set. In reviewing the data gathered, we determine that it was necessary to fill data gaps so that the extracted rows matched the format of the BSEEPublic dataset as closely as possible. There were several columns that needed to be created based on data in the free-form text in the incident summary column. For example, we identified several incident reports with description text that could be used to label incident records in accordance with our categorization taxonomy. Even though his task could be done manually by human analysts, machine learning methods offer ways to increase efficiency and scalability.
4) Populating a data model. This phase of the research project was approached as a proof of concept to demonstrate the potential value of the data model, data gathering and data transformation steps when implemented as an operational solution. We constructed the proof of concept using Microsoft SQL Server along with Microsoft Power BI for analytics and visualization, which involved three steps; Create database, load data, build analytics. The initial steps of the proof-of-concept development process
created the structure to load the data into. We developed the database architecture to directly reflect the data model results. After we loaded the data into the database, we created database views. Although the data in this proof-of-concept have been reduced too far to generate true results, the interface

immediately highlighted interesting trends and findings. For example:

- Fixed facilities have the highest rate of well activity incidents per Bbbl produced but the lowest incident rate per well and per well activity.
- TLPs have a disproportionate number of well activities per well.
- Fixed facilities with high incident rates tend to have large numbers of wells.
- It is very important to realize that these are only notional conclusions; the data are too incomplete for any true risk analysis results. However, despite the limited data, the data model and associated analytics proved insightful.

Communications, Outreach, and Dissemination Activities of Project

ABS issued a press release related to the award of this project. Article can be found at: https://ww2.eagle.org/en/news/press-room/abs-awarded-two-point-one-million-dollars-for-safetyresearch.Html

Dr. Wang's team gave a presentation at the Society for Risk Analysis Conference; Title: Data Science Tools for Aggregation and Analysis of Safety Incident Data

Date: 12/9/2021