PROGRESS REPORT

1. **Project Title:** Information Retrieval via Knowledge Graphs Developed for Aircraft Accidents Database and Aircraft Manuals.	DST No: IMP/2018/001627/IT
2. PI(Name & Address): Dr. Pushpak Bhattacharyya, Director and Professor, IIT Patna, Department of Computer Science And Engineering. Email:pushpakbh@gmail.com	Date of Birth :July 03, 1962
3. Co-PI(Name & Address):	Date of Births:
3.1 Collaborators (Name and Address)	Date of Birth:

4. Broad area of Research :Artificial Intelligence
Sub Area: Knowledge Based System

5. Approved Objectives of the Proposal:
To develop a Knowledge Graph for Aircraft Accident and Aircraft Manuals to retrieve useful information through it.

Date of Start: July, 2019	Total cost of Project: 86,40000
Date of completion July, 2022 (3 years)	Expenditure upto March 31, 2020 Capital: 645350/-; General: Rs. 276,000/-

6. Methodology

The research method employs collecting the open data sets, safety reports related to Aircraft Accidents and Aircraft/Subsystem Maintenance Manuals and identify entities and relationships between the two. The NTSB Aircraft Accident Report is used for the development of Aircraft Accident Knowledge Graph. The NLP techniques are used in the generation of knowledge graphs. The Knowledge Graph is evaluated by querying it using the queries from various domains of Aircraft Accidents, provided by Honeywell. A Graphical User Interface is designed to make it easier to query the Knowledge Graph in a Natural Language. Following are the various development phases of the project.

- **Knowledge Graph Development:** The first Phase of the project is building of Knowledge Graphs forthe Aircraft Accident Report Domain. The Knowledge is developed by processing the unstructured data available for the Aircraft Accident at NTSB site. The processing of unstructured data involves the steps like Entity extraction, Relation Extraction, Sentence Clustering,nomenclature of different clusters and then ontology creation.Finally the ontology is populated with the extracted data from NTSB.
- **User Interface:** The query language used in ontology is sparql that is a structured language. The querying is made user friendly by developing a Natural language query system. The tool is developed using spacy library and owlready2 driver in python to convert Natural language query into sparql query. The key outcome of the project is the answering of the Aircraft accidets related query. The Figure 2 below shows the snapshot of query asked by user in natural language that is 'Which accidents have substantial damage?' and the output is shown by the interface after retrieval from the knowledge graph.
- **Accident Cause Extraction:** The cause identification of Aircraft Accident is one of the challenging tasks. The cause analysis of Accident from NTSB data is extracted that is in unstructured format.

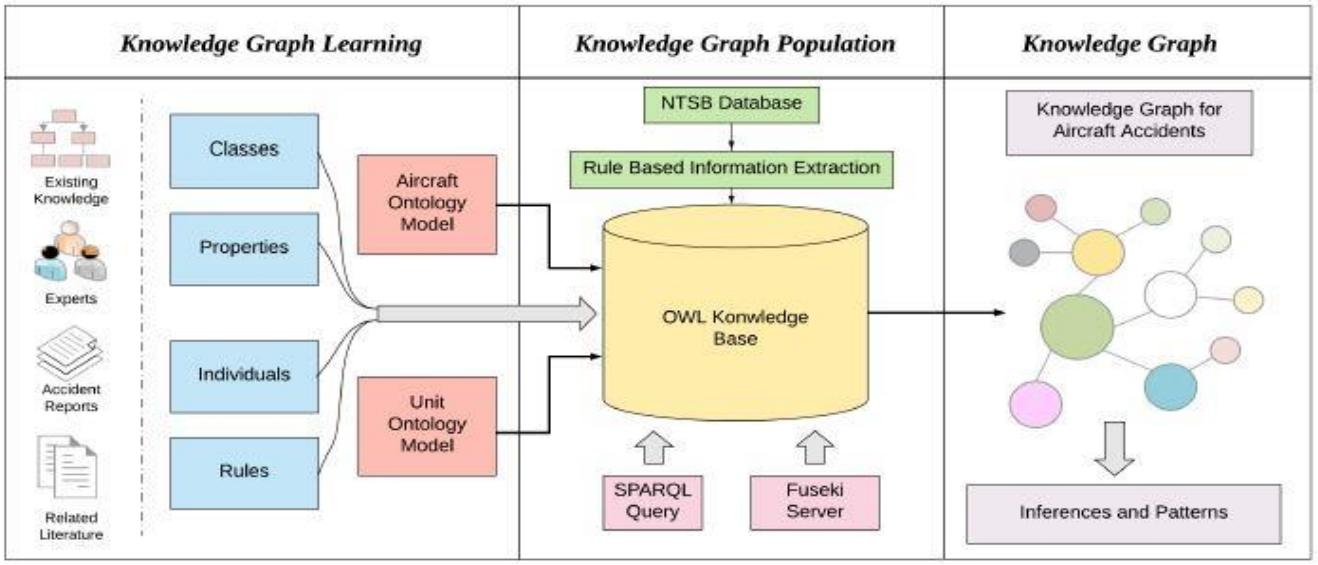

Figure 1 : Knowledge Graph Construction Process

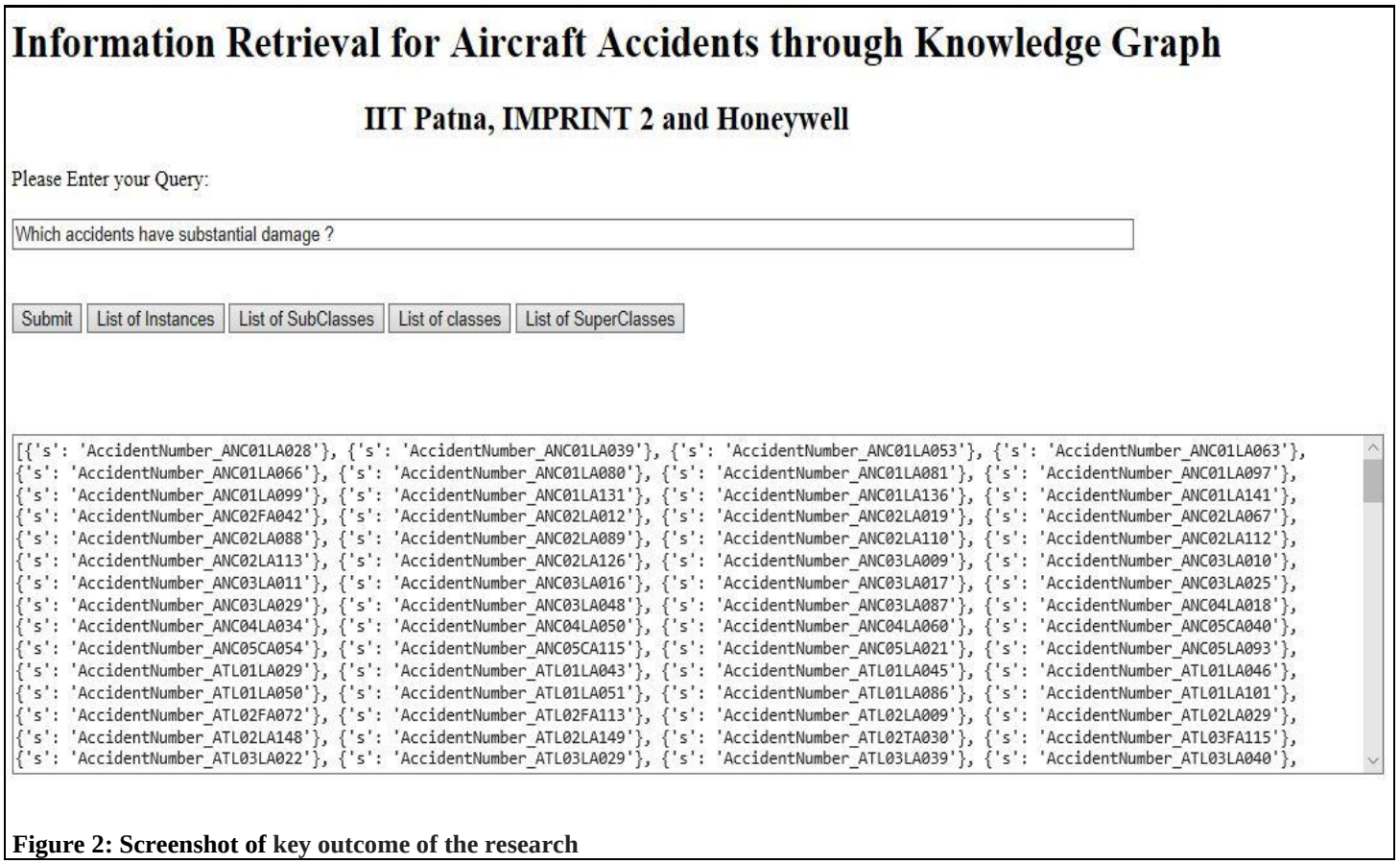

Figure 2: Screenshot of key outcome of the research

7. Salient Research Achievements:

7.1 Summary of Progress

The work done till now on the project is summarized below:

1. **Motivation:**It is started by describing the motivation behind the project and explained why the aircraft safety is of utmost importance. Then the problem statement of the project is observed and also explored the existing aircraft accident databases maintained by different investigating organizations. We introduced the concept of ontology and knowledge graph and also explored existing ones. After surveying the existing knowledge graphs, the benefits and the problems of the knowledge graphs are discussed. The knowledge representation language and query language (SPARQL) is studied and at last, saw the knowledge engineering software such as Protégé.
2. **Use of Ontology/Taxonomy:** The advantages of using ontology are-
    - Ontology has open world assumption
    - It Can represent semantic relationships present in the data
    - Ability to generate automated reasoning
    - Additional information can be inferred using present information.
    - Knowledge Graphs are easily scalable
3. **Data Collection:** The Aircraft accident data is collected from National Transportation Safety Board (NTSB) site having freely available data. We selected the NTSB database for our experimentation. The key reasons for selecting this database are as follows:
    - NTSB stores investigation reports of civil aviation accidents and the scope of the project is limited to enhancing the safety of civil aviation.
    - NTSB has thousands of investigation reports and other organizations have less number of reports.
    - NTSB investigation reports follow a consistent format which is relatively easy to process and extract information.
4. **Knowledge Graph Development:** Studied the overall knowledge graph learning process and explained the importance and

steps of data pre-processing like cleaning, annotating and lemmatization. Discussed the entity extraction processes and then moved to the relation extraction methods. Learned about the unit ontology and discussed the motivation behind adding of the unit ontology. The steps followed for Knowledge Graph development are:

- **Pre-Processing of Data:**To get good results from an algorithm, the data fed to that algorithm must be clean and should not contain any noise. The steps of data pre-processing are cleaning, part-of-speech tagging and lemmatization.
- **Information Extraction:**The information in the form of entities and relations are extracted from the NTSB data. The data is extracted from the NTSB reports in an unstructured format. Discussed about the feedback from the knowledge graph experts and domain experts in designing of aircraft accident knowledge graph.
- **Unit Ontology:** The unit ontology is used for the quantities with units. For example distance quantity, angle quantity, weight quantity, velocity quantity etc. Then illustrated the existing ontology and studied the integration process of unit ontology with our knowledge graph. If we do not include the unit ontology, there could be no way to fetch unit of a quantity or retrieve that quantity in some different units.
- **Ontology Creation:**The structure of ontology is created by the clustering of the NTSB data and then analysing these clusters by domain experts to obtain the taxonomy.
- **Knowledge Graph Population:** It is the task to populate the data from XLSX to the knowledge graph. Cellfie software is used to populate the knowledge graph. Cellfie software takes the XLSX file and JSON rules as input and populates the graph based on the rules provided.

5. **Knowledge Graph Evaluation:** The Knowledge Graph evaluation is performed with the help of evaluation from Ontology/KG expert and also from domain experts. Then studied existing ontology for aircraft safety manuals. At last, saw the approach of designing the knowledge graph. Created a list of test queries from the different domains of the aircraft accident. The domains for querying are provided by the domain experts from the Honeywell. The knowledge graph is tested with the help of these test queries.
6. **Interface Designing for Natural Language Query:** A tool is developed to convert the Natural Language query into Sparql query. A rule based approach is used for the conversion of Natural Language query/keywords into sparql. The conversion is done by doing the POS tagging of the query and then matching it with ontology triples to extract the closest triple. Implemented the owlready2 driver to access the Knowledge base directly from the python platform. Here, the query is asked by user in Natural Language format and the system directly gives the output from the aircraft accident Knowledge graph. The Graphical User Interface provides a user friendly platform for querying the Knowledge base in natural language or using relevant keywords. Here, the Django framework in python is used for the development of this user interface.
7. **Accident Cause Extraction and Analysis:** NTSB data is analyzed and then the cause behind aircraft accident is extracted from the NTSB data. A number of cause events are involved behind an aircraft accident and it is a challenging task to find the sequence of causes. The correlation analysis is to be done to find the cause event sequence of accident causes.

7.2 New Observations:

- It is observed that the knowledge base can be improved as an intelligent system by applying some semantic rules in it.
- The application of knowledge graph may be used in the field of safety enhancement of the aviation domain.
- The application can also be used in the development of an intelligent system that could automatically take a precautionary action before any mishap.
- The cause analysis of recurrent accident**/incidents** would help in recognizing patterns that may help in **understanding limitations in system design , crew or maintenance procedures.** There would be a need to develop a time plot of recurring risks with information from accident/incident reports.

7.3 Innovations:

- It provides a platform where any non technical person can ask the question related to aircraft accident in a Natural language to get the answer from the knowledge base.
- The unstructured information of NTSB is converted into a structured format that enables us to easily retrieve/analyze the specific information.
- The Graphical User Interface provides a user friendly platform to perform a query operation.

7.4 Application Potential:

7.4.1 Long Term

- The application can be used in a combined way for both Aircraft Accident and Aircraft Safety Manuals that can help in retrieving some of the useful hidden information.
- The Application may provide the accident information based upon the various patterns of the cause events.

7.4.2 Immediate

- The Application can retrieve some basic information regarding aircraft accident like type of accident, engine, aircraft, etc.
- The query can be asked in a natural language or by using some relevant keywords.

7.5 Any other

The approach used in designing of interface for Natural language Query may be useful for other domains also

Research work which remains to be done under the project(for on-going projects)

(i.)Generation of the Event Sequencesby the cause analysis of the Accident.
(ii.)The development of the Knowledge Graph for the Aircraft Safety Manuals.
(iii.)Merging of the Aircraft Accident and Aircraft Safety Knowledge Graphs.
(iv.)Implement methods to make knowledge graph infer additional safety measures.
(iv.)Optimization, Pruning, and Adaptation to new data sets for the knowledge graph should be done.

PhDs Produced no: Ongoing-	Technical Personnel trained: Qumar Ibrahim, (PhD Scholar, IIT Patna)	Research Publications arising out of the present project:

Patents filed/to be filed: NA

Major Equipment(Model and Make)

S No	Sanctioned	Procured (Yes/No), Model & make	Cost (Rs in lakhs)	Working (Yes/No)	Utilization
1	GPU based Server	**Yes,** GPU Card NVIDIA 2O8O Ti 11 GB Blower Make: PNY	6,02,237.00	Yes	100