SECTION 'C': DETAILED RESEARCH PROPOSAL

The information given in this section must be adequate and sufficiently self-contained to enable the advisory committees and referees to assess the merits of the scheme. In particular, it must furnish the following information:

(a) **Title of the project:** Leveraging Domain Ontology and Knowledge Graphs for Efficient Inference from Geospatial Big Data through Machine Learning

(b) **Objectives:**

1. To represent the geospatial big data using appropriate knowledge graphs
2. To design machine learning methods for imposing inference rules on knowledge graphs
3. To extract most reliable details from knowledge graphs using inference rules
4. To devise efficient algorithms of inference from geospatial big data
5. To design a web based application for processing the devised algorithms

(c) Background, origin & relevance of the proposed R&D work to state priorities

The present era belongs to data, especially big data, which is being produced at a huge amount per day. Traditionally, the term data refers to raw facts and figures which tend to become big data when the data is not only big in size but also big in production and processing. The features of big data can be characterized by 5 Vs which refer to volume, velocity, variety, veracity and value (Jin et al. 2015, Liu et al. 2020). Further, big data can be divided into 2 broad categories, viz. data from the physical world and data from human society (Jin et al. 2015). The data from the physical world can be measured through sensors, observations and experiments whereas the data from human society can be acquired from a variety of sources including social media.

Irrespective of the sources of data, the big data inherently includes spatial information, i.e., the location where the data is generated. With reference to the Earth’s surface and its surrounding area, the spatial data is termed as geospatial data. Geospatial data has been widely used in Geographic Information Systems (GIS) for various applications, e.g., disaster management and response, agriculture risk management, environmental planning, and water resource protection (Zhang et al. 2021). Accordingly, processing the geospatial data is of crucial importance for the research perspective in the field of GIS. As the amount of geospatial data being generated per day is so large, its storage, processing and retrieval have become a challenge for the researchers. Since geospatial big data is concerned with the geolocation of the data, both the data from the physical world and the data from human society are to be considered as geospatial big data.

As far as the data sharing is concerned, numerous platforms are now available including USGS Geo Data Portal, Open Data Portal, United Nations Digital Library, NASA Common Metadata Repository (CMR), ISRO Bhoonidhi. Nevertheless, enabling intelligent and efficient spatial data sharing and communication among various users is a tremendous task. The major challenge is building meaningful semantics between the available data products using spatiotemporal similarity measures so that the users may be able to find appropriate details. Data visualization is another problem while dealing with geospatial big data which needs a lot of computational power and skills. The representation of geospatial big data efficiently, e.g., using knowledge graphs may help the users to visualize and analyze the details in a different domain. In this direction, efficient algorithms may be devised.

Satellite Data Status

Although the geospatial data includes various types of datasets, the current project will mostly focus on the satellite based data including the ground truth. ISRO provides a variety of satellite data which is available for various applications including weather forecasting, monitoring etc. It provides satellite imagery required for the developmental and security needs of the country. Further, it provides satellite imagery and specific products and services required for application of space science and technology for developmental purposes to the Central Government, State Governments, Quasi Governmental Organisations, NGOs and the private sector. ISRO has provided the users Bhuvan and Bhoonidhi portals for the access and procurement of a variety of satellite data. Bhoonidhi enables access to an extensive archive of Remote Sensing data from 46 satellites, including Indian and Foreign Remote Sensing sensors acquired over 33 years (https://www.nrsc.gov.in/sites/default/files/pdf/ebooks/UIM-2022/uim_14.pdf). The portal also provides regional data of Landsat and Sentinel satellites in India.

The variety of satellite data available from Bhoonidhi portal include Aqua, Cartosat-1/2/2A/2B/2C/2D/2E/2F/3, GSAT29, HYSIS, IRS-1A/1B/1C/1D, KompSat-3/3A, Landsat-5/8, MicroSat-1, NOAA-11/12/14/16/17/18/19, Novasar-1, OceanSat-1/2, RISAT-1/2B/2B1/2B2, ResourceSat-1/2/2A, ScatSat-1, Sentinel-1A/1B/2A/2B and Terra. As per the policy, the dataset with resolution coarser than 5m is free and open for all whereas the dataset with resolution between 5m and 0.5m is free for Government agencies for R&D applications and priced for private agencies. The Bhoonidhi Vista provides full resolution data visualization on an online map whereas Bhoonidhi Upagrah provides the live tracking of satellites as well as prediction of the position of the satellite at a future time.

Human Activity based Data Status

Similarly, human activity based big data is available from a variety of sources. Few of them are listed below:

http://iot.ee.surrey.ac.uk:8080/datasets.html#traffic
http://iot.ee.surrey.ac.uk:8080/datasets.html#parking
http://iot.ee.surrey.ac.uk:8080/datasets.html#pollution
http://data.surrey.ca/dataset/water-meters/resource/99fe8786-6329-49f7-ae92-2c3b8f6e4778
https://www.kaggle.com/datasets/programmerrdai/urbanization
https://www.kaggle.com/datasets/thedevastator/global-urban-area-indicators
https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india
https://www.kaggle.com/datasets/fedesoriano/air-quality-data-in-india
https://www.linkedin.com/pulse/data-science-architecture-urban-planning-riddhi-sarda/

(d) Practical/Scientific utility:

The project will deal with various types of geospatial big data and their representation methods. Since geospatial data is represented as raster, vector or attributed data, a variety of representation methods may be devised for each of these data sets. Further, big data have their own challenges of representation which make geospatial big data more complex to represent. Among the traditional methods of representation including spatial data files and attribute data files, semantic networks may be followed for representation which utilizes the semantic rules in a graphical manner. Accordingly, the concept of knowledge graph evolved which not only represents the spatial details but also provides effective rules to extract the details from the complex data storage. Traditionally, the retrieval methods include SQL based query which are extended with XML based representation and retrieval. The methods are extended with SPARQL while knowledge graphs are used for representation thus making the retrieval process effective. Once the complex geospatial data is represented efficiently, the challenges of data retrieval and inferring information from the complex data sets are at hand. Traditional methods like regression extended to logistic regression and probabilistic models are frequently used for predictions. The Markov model is also a popular method in the geospatial domain which may be improved with the aid of machine learning in future. The retrieval of information from complex geospatial big data is useful for various applications like disaster mapping and mitigation, urban planning, agriculture mapping and crop produce estimation. The human activity based spatial data is of importance for various applications where the spatial footprint of an individual person may be utilized. For example, the spatial footprint may track the activity of an individual person leading to the details related to the local activities in a specific geographic region.

The goal of this project is to increase the effectiveness of geospatial big data inference by combining domain ontologies, knowledge graphs, and machine learning strategies. Large amounts of data about urban settings, transportation infrastructure, and other geographical factors can be found in geospatial big data, making it difficult to process and derive useful insights from it. The project suggests creating domain ontology specifically for geospatial big data that will capture the essential ideas, connections, and characteristics of the urban environment in order to address this problem. The ontology serves as a formal representation of knowledge, fostering communication and understanding among the various parties involved in the analysis and decision-making of geospatial data. The subsequent steps entail building a knowledge graph that incorporates numerous data sources, including geospatial data, demographic data, infrastructure data, and environmental data. A unified representation of the interconnected entities, attributes, and relationships is created by the knowledge graph, allowing for flexible querying and reasoning. It is possible to create decision support systems that use machine learning algorithms to produce simulations and recommendations for assessing the potential effects of various policies or interventions in urban planning.

The project seeks to make inference from geospatial big data more effective and accurate by utilizing the integration of domain ontology, knowledge graphs, and machine learning. The present proposal will deal with the spatial data for inference from complex data sets for urbanization applications as a case study. It will infer the urban growth in future years and urban state on the basis of available geospatial details.

(e) Up-to-date Review of research conducted/being conducted on the subject in India and abroad:

(i) At the sponsoring institution (State preliminary work already done, techniques standardized, data collected etc.).

- In recent years, both PI and Co-PI have worked on various projects related to geospatial data analysis and big data analytics with publication in related fields.
- The PI has more than 18 years of research experience in the field of satellite image processing and analysis. The recent development of PI

includes the funding by DST to organize a 21 days Summer School on Geospatial Science and Technology which was conducted successfully during 23rd May and 12th June 2022.

- The Co-PI has more than 20 years of experience in the field of Machine Learning application developments. Co PI has published many papers in the area of Machine Learning applications especially target to support geo spatial data analysis. She has published papers in the area of uncontrolled urbanization.

(ii) Research work done and in progress in India.

It is critical to keep in mind that as new developments and initiatives are launched, the state of research in the country may change over time. An active and quickly developing area of research is handling of geospatial big data using domain ontologies and knowledge graphs (D'Aniello et al. 2020). In recent years, there has been a lot of interest in using geospatial big data, domain ontologies, and knowledge graphs to generate inference. In a variety of fields, such as urban planning, disaster management, transportation, and environmental monitoring, these technologies have the potential to offer insightful information and support decision-making processes (Chandra et al. 2022, 2023). Dev at al. (2022) have designed a geospatial database for implementing enhanced local governance with spatial analysis for Manesar area, Gurugram, Haryana. The web-based GUI developed by the team using MicroStation and MS Access is able to view and update attributes of the features in real time. The research demands project-oriented creation and customization of spatially responsive SQL. Doad et al. (2022) present a spatial multiple criteria evaluation analysis (SMCE) using GIS and remote sensing techniques with spatial datasets to identify the geo-environmental values for a watershed in central India. A Multi-Criteria Decision Analysis (MCDA) technique is implemented by the researchers on spatial databases for identifying relevant natural and human influenced factors which have a direct bearing on the environmental situation in an area.

Further, by exploring the recent available literature, research publications, and conferences in this area, some of the important research areas and methods include Geospatial Data Integration, development of Domain Ontologies from national perspectives, construction of knowledge graph, reasoning and query processing using knowledge graphs (Barik et al. 2022, Marcinko 2022, Mir et al. 2022, Prathap 2022). Developing decision support systems that leverage the knowledge graph and ontologies to provide real-time recommendations, optimize resource allocation, and simulate the impact of different interventions or policies. It integrates methods from geospatial analysis, semantic web, machine learning, and data management to enable effective decision-making in intricate national contexts (Chandra et al. 2022). The national geospatial program is highly supporting more research activities in this direction.

(iii) Research work done and in progress abroad.

Internationally, research on applying domain ontologies and knowledge graphs to manage geospatial big data has attracted a lot of attention and is still advancing. Worldwide, a large number of nations and academic institutions are actively investigating and advancing this field of study. In terms of research into knowledge graphs and geospatial data analytics, the United States has been at the forefront (https://www.geospatialworld.net/consulting/gki-phase-2/gw-assets/pdf/GKI-Report.pdf). Europe has concentrated on constructing knowledge graphs with geospatial data and domain-related ontologies. The present state of knowledge in this field has significantly advanced due to the contributions of European research institutions and industry partners Universities like the University of Southampton, Imperial College London, and the University of Oxford are also doing research in this field (https://www.geospatialworld.net/consulting/gki-phase-2/gw-assets/pdf/GKI-Report.pdf). Other nations with research projects centered on geospatial inference using domain ontologies and knowledge graphs include Australia, Brazil, Japan, South Korea, Brazil, and India. The use of domain ontologies and knowledge graphs in geospatial inference has advanced by means of initiatives funded by agencies like NASA, NGA, and the National Science Foundation (NSF). Research publications, conferences, workshops, and international collaborations have all contributed to advancements in this area. When using domain ontology and knowledge graphs to address common issues in geospatial inference from big data, researchers from various nations frequently work together on projects and share insights (Chandra et al. 2021, Tiwari et al. 2022, Zhang et al. 2012).

With the advent of big data, a significant increase in heterogeneous data, especially in the geospatial domain is observed. Since heterogeneous data with complex spatial relationships may lead to information disorientation and overload, it is necessary to deal with such data carefully. A heterogeneous retrieval method based on knowledge graph is proposed by Liu et al. (2021) which have 3 advantages, viz., the semantic knowledge of geospatial data is considered, more detailed information can be obtained and the retrieval speed can be improved. The method also utilizes the combination of knowledge graphs and GIS to acquire related geospatial information intelligently with the case study of Zhengzhou railway station, China. Further, a multisource open geospatial big data fusion has been done by Priyashani et al. (2023) for demarcation of urban agglomeration in Colombo, Sri Lanka. The study used a variety of data sources including Google Map API, Twitter API, NASA Earth service data, Google Earth Pro and Open street map which not only deals with the heterogeneous data but also with a complex set of geospatial big data.

(f) Actual plan of work:

(This is an informative statement for scrutiny by the scientific panel indicating essential phases/ items of contemplated programme giving insight into the methodology and experimental techniques to be employed for executing the research programme plan. A well prepared year wise plan of work is essential for the appraisal of research proposal).

(i) The main items of observations to be recorded.

(ii) Arrangement for analyzing data and the name and designation of the statistician associated in the programme planning.

(iii) The items of investigation for which collaboration indicating the collaborators with other Section/Department/Institute as in proposal.

(iv) Field visits for dissemination of scientific outcome of scientific value.

(v) Flow chart of work plan/methodology.

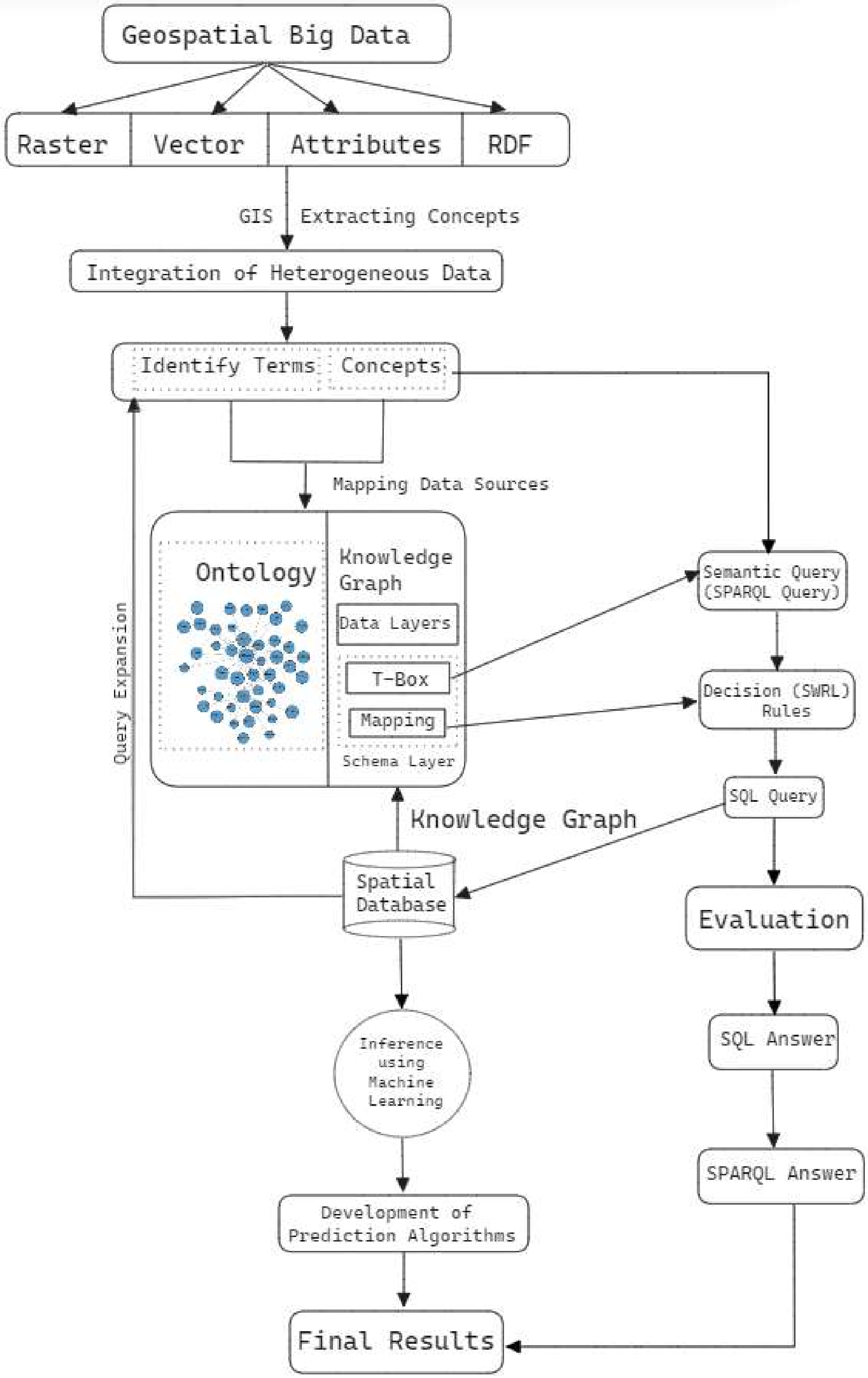

Figure 1. The flowchart of the proposed methodology

The methodology includes representation of geospatial big data using domain ontology and knowledge graphs followed by information extraction using machine learning methods. The heterogeneous data will be integrated together to create the knowledge graphs. The algorithms for inference from geospatial data will be developed for efficient extraction of details from the represented data. The semantic rules based on ontology will help to design the inferencing appropriately.

(vi) Deliverables of the proposed work.

Objectives	**Activity**	**Start month**	**End month**	**Deliverable/Output**
Objective 1	Data Collection including field data, Representation of data in the form of database and knowledge graph	Month 1	Month 12	Knowledge graphs and Geospatial Big Database, Annual Progress Report
Objective 2	Generation of inference rules using machine learning tools, Refinement of the algorithms	Month 4	Month 30	Machine learning algorithms for inference from geospatial big data, Publication(s) in form of report/research article for the task done, Annual Progress Report
Objective 3	Extraction of reliable information from knowledge graphs, Prediction of the geographic expansion of urban region within the study area	Month 7	Month 30	Reliable information extraction methods from knowledge graphs, Publication(s) in form of report/research article for the task done, Annual Progress Report

Objective 4	Designing efficient algorithms based on machine learning for inference from knowledge graphs	Month 24	Month 32	Algorithms of inference from geospatial big data, Publication(s) in form of report/research article for the task done, Annual Progress Report
Objective 5	To design a web based application on the basis of inferencing from knowledge graphs	Month 25	Month 36	A web based application of inferencing from geospatial big data, Publication(s) in form of research article, Final Project Report

Table 1. Time line and deliverables of the project

(g) Likely deliverables/ expected outcome of the proposed work

The project is meant for fulfilling the goals of the Geospatial Program of the Government of India which will produce a web based tool for processing the devised algorithms for inference from geospatial big data. Besides the web based application, project reports and research articles based on the work done will be the outcome of the project. These outcomes will be helpful for spreading the geospatial knowledge throughout the community and promote the motive of the national geospatial program.

(j) Parameters for monitoring the progress

The progress of the project will be monitored every 6 months in form of a presentations followed by the progress report submission. The outcome and deliverables will be evaluated accordingly.

(k) Suggested post project activities

Promotion of the prototype developed and the deliverables.

(l) Beneficiary groups/departments

a. ISRO
b. DRDO
c. CSTUP/DST
d. Academic Institutions