**Energy AI and Knowledge Graph – Concept Note**

UNDP Sustainable Energy Hub

October 11, 2023

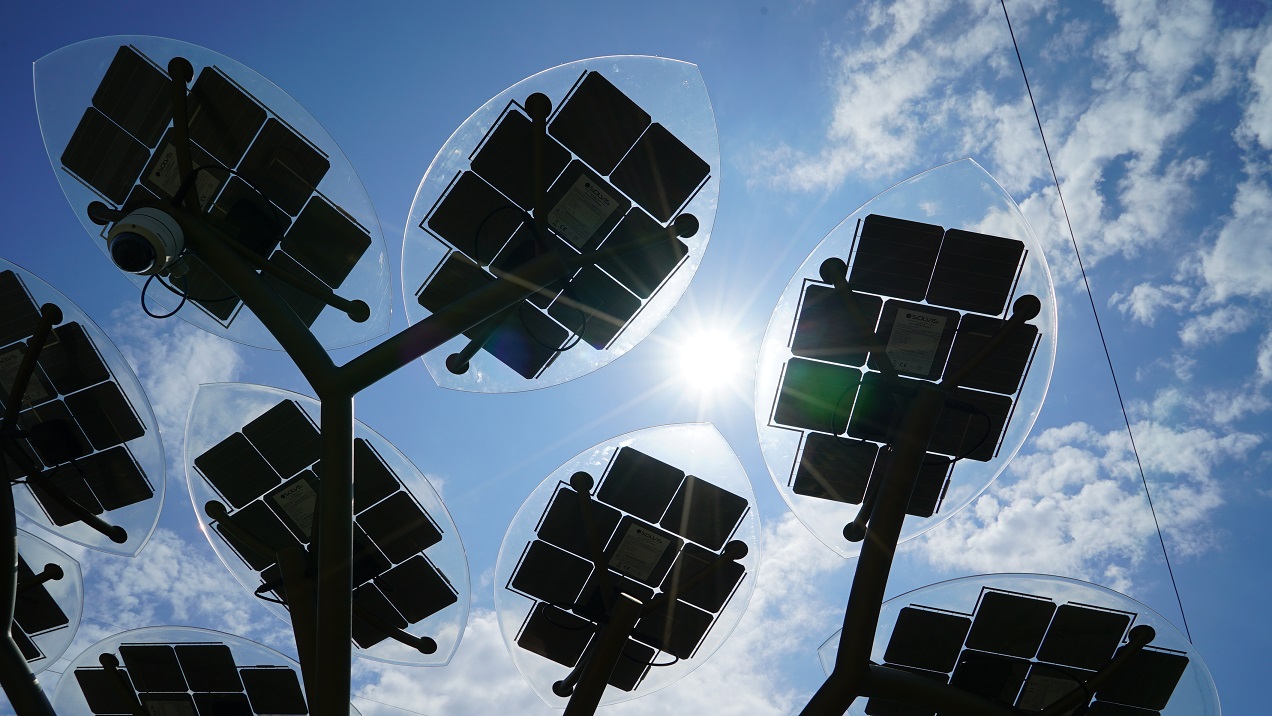


Table of Contents

[I. Introduction 1](#_Toc42057514)

[II. Approach 2](#_Toc1224890513)

[a. System Architecture 2](#_Toc939119921)

[b. AI Response Engine 2](#_Toc891335699)

[c. User Interface 4](#_Toc1940992650)

[d. The Energy Knowledge Graph 5](#_Toc40217331)

[III. Implementation Plan 6](#_Toc1689937732)

[A. Project Timeline 6](#_Toc821193892)

[B. Personnel and Budget 7](#_Toc583223718)

[D. Next Steps 9](#_Toc241925373)

[IV. Appendix 10](#_Toc37359327)

# Introduction

In the pursuit of sustainable energy solutions and informed decision-making, the Sustainable Energy AI in development by the UNDP Sustainable Energy Hub combines recent advances in LLMs with a robust energy knowledge graph to create a powerful new tool to support access to insights on energy. The Energy AI synthesizes insights from a comprehensive database of compiled publications, national policies, real-time APIs, and a database of indicators. This approach can provide access to knowledge and intelligence which can empower policymakers, researchers, and development practitioners with accurate, transparent, and contextually rich information for technical capacity-building and data-driven decision-making.

The AI will be built using an ensemble model with access to a compilation of resources and datasets, rather than an LLM only approach. This approach provide advantages and limitations. Rather than a single language model that can provide text responses directly to queries, this approach involves more steps to first identify the most relevant documents and datasets within the database, directly access this information, augment it with insights from the curated knowledge graph, and then compile an answer from these results. This allows a more targeted approach to respond to certain user needs and provide answers directly linked to the original literature, which increases transparency and ability to cite sources and allows a more mechanistic approach to building a model to respond to a very specific set of types of questions for certain stakeholders.

Key advantages of this approach over using ChatGPT alone include:

* **Citation of Sources:** This approach allows for proper citation of sources, ensuring the credibility and traceability of information.
* **Transparency:** The Knowledge Graph (KG) is 'transparent' rather than 'opaque' like ChatGPT, offering complete control over parameterization, architecture, sources, etc.
* **Minimized AI Hallucination:** All information is derived directly from a meticulously reviewed knowledge graph, significantly reducing the possibility of AI-generated hallucinations.
* **Real-time Integration:** An ensemble model enables direct connections with APIs and real-time data sources, ensuring the most current and relevant information is always accessible.
* **Chart generation:** Since there is access directly to the datasets, we can generate graphs based on data mentioned by the user or related to the response.

# Approach

The approach will involve a more mechanistic ensemble model than traditional LLMs. This modular approach will allow the system to be constructed based on a set of verified modules that function according to the specific user needs and system requirements designed around energy stakeholders and the use case of sustainable energy thought leadership.

## AI Response Engine

A black board with text and colorful text

Description automatically generated The proposed architecture has four steps to generate a response from a user query:

Data Integration, User Query Processing, Synthesis, and Validation.

1. User Query Processor:

The user query is analyzed and the entities are extracted.

* + 1. Identify the relevant documents and execute semantic search to extract text segments most relevant for the query.
    2. Identify the dataset and filter parameters. If a data point is being retrieved, access it directly. If analytics are required, the query processer will engineer a query for PandasAI for this dataset, and run the PandasAI API call to evaluate statistics or generate a graph. PandasAI is to be replaced with a custom data analytics module when the scope allows.
    3. From the Knowledge Graph, retrieve summaries of each entity as well as the category and description of all relations between each pair of entities

1. Data Integration

The data integration layer brings together content from documents and datasets as identified in the query processer, using semantic search and data filtering.

* + 1. Documents
    - **National Policies and Plans –** A compilation of over 600 national documents from the 171 countries UNDP has a presence in **(**Plans on energy transition,roadmaps, investment plans, efficiency plans, NDCs, Country Programme documents)
    - **Energy Publications (**UN agencies, IEA, World Bank, etc.)
    - **Journal Articles**
    - **SEH Regional Strategies and Policy Briefs**
    1. Datasets
    - **Indicator Database (**World Bank, IEA, IRENA, etc.)
    - **Moonshot Tracker** - UNDP Portfolio on energy-related activities (internal\_
    - GIS Data - Global gridded datasets related to energy nexus
    - **Real-time APIs -** Live data on energy production, consumption, and environmental metrics. This data requires an additional cleaning and processing layer to prepare it for processing based on the defined data standard.

1. Synthesis

Write the prompt engineering instructions and use ChatGPT to synthesize an answer based on a compilation from:

* original user query
* excerpts from the most relevant documents
* relevant data points and real-time information
* references and citations for all text and data
* information for all entities and relations from knowledge graph

1. Validation

A final layer to review output for sensitive topics and information, data accuracy and quality, validity of citations and source, etc.

## User Interface

1. Web app

The interactions will all take place within the Sustainable Energy Knowledge Platform, a tool currently in development by the SEH. The primary interface will include several components, but at its foundation it will work in a similar fashion to a text interface such as ChatGPT with several additional layers of enhanced functionality.

1. Knowledge Graph visualization engine -

Two primary versions of the knowledge graph will be generated: one will be designed as a data product for enhancing the results of the Energy AI, and one will be a more minimal collection of entities and nodes that are designed to be navigated by development actors in an interactive graph visualization tool. This visualization will be built using D3 and will be a responsive web visualization tool to explore the network graph and provide access to the base of knowledge matching the entities contained within user queries.

1. Embeddable chatbot

The tool will be embeddable within a minimized interface to be integrated within the other digital tools in development by UNDP on energy.

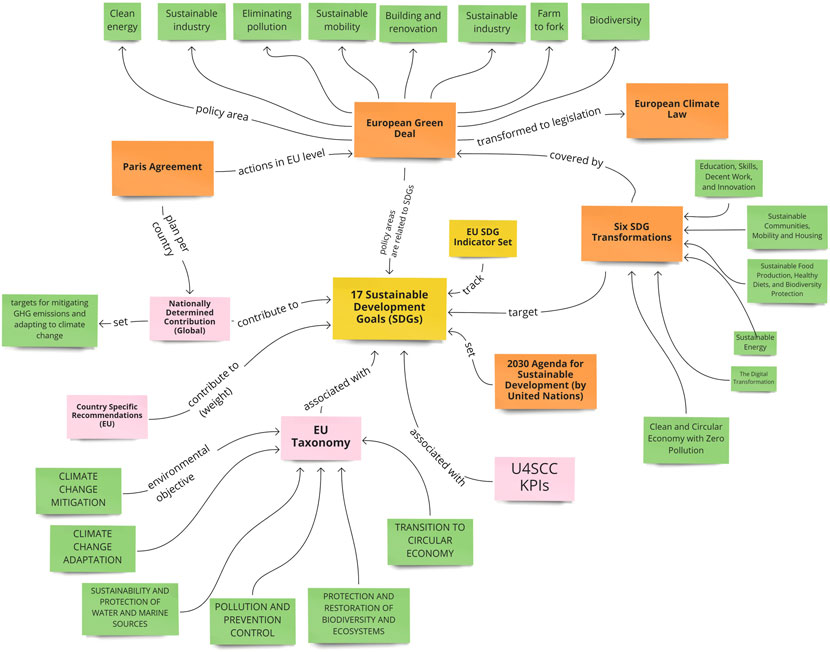
1. Query suggestion engine

An interface will be designed around the query suggestions, which will be critical to aid a user to be able to explore the functionalities of the AI and to be able to discover working queries. Because the set of queries that provide meaningful responses is more limited than a general LLM, this feature will be key for users to get the most out of the tool. The interface will be reactive to users responses when they start to type, and in case a user submits a query that is not supported, it will provide helpful suggestions for better queries more aligned with the architecture of the AI.

* 1. **Document Viewer:** Allows users to explore compiled publications, national policies, and other documents.

## The Energy Knowledge Graph

A knowledge graph is composed fundamentally of **Entities** and **Relations**. Each entity and relation contain type and descriptions (100-1000 words), generated initially with AI and then reviewed by experts. The AI will be trained on the compiled set of several thousand publications compiled on sustainable energy and national policy and planning documents.



*Integration of the Sustainable Energy Academy into the Knowledge Graph*

Material from the 16 modules will be used as the initial training material for the knowledge graph. This immediate base of knowledge will be used to validate other data as it is ingested from other documents. This integration will also allow a new way of interacting with the material from the Sustainable Energy Academy through a dynamic and customized interface based on user needs.

# Implementation Plan

## Project Timeline

**Phase 1: Project Initiation (1 month)**

* + 1. Conduct research on energy knowledge graphs and user requirements.
    2. Finalize project scope and workplan.

**Phase 2: Knowledge Graph Development (2 months)**

* + 1. Collect and organize data from diverse sources.
    2. Aggregate data into relevant categories for the knowledge graph.
    3. Design schema, develop algorithms, and implement the knowledge graph.

**Phase 3: AI Model Development and Training (2 months)**

* + 1. Choose suitable LLMs and configure them for natural language processing.
    2. Train LLMs using the curated knowledge graph data.
    3. Integrate AI models with the knowledge graph API.

**Phase 4: User Interface Development (1 month)**

* + 1. Design user interface and create interactive prototypes.
    2. Develop frontend and integrate it with the backend modules.

## Personnel and Budget

1. Required Budget -

The minimum required personnel costs for implementing this proposal include:

|  |  |  |
| --- | --- | --- |
| Title | Budget | Minimum Scope |
| Project Manager/ Designer/ Data Scientist | $20,000 | Develop and manage workplans, design and implement system architecture, design and lead development of interfaces, coordinate data compilation and cleaning, draft documentation and project reports  Deliverables: MVP |
| NLP Consultants | $12,000 | Research, recommend, and implement NLP models for energy-related queries. Develop autocomplete and prompt suggestion features.  Deliverables: Integrated NLP system with query processor, synthesis engine, data retrieval module |
| Expert Reviewers for Knowledge Graph | $10,000 | Validate knowledge graph content, provide domain expertise for revision.  **Deliverables:** Enhanced knowledge graph |
| Web Developers | $8,000 | Develop interactive web and mobile interfaces, integrate frontend with backend modules, and develop charts and knowledge graph visualization engine  **Deliverables:** Web app |

|  |  |  |
| --- | --- | --- |
| **Total:** | **$50,000** |  |

API Credits (including using ChatGPT through Azure Web Services) will be an additional ongoing cost, but we can use general UNDP credits and should remain fairly low unless additional services are identified.

**Existing Team**

The Sustainable Energy Hub team and partners working currently on this project:

1. Project Management, Solution Architect, Design, Data Science– Ben Keller (SEH Consultant)
2. User Query Processing and Responses Synthesis – NLP Consultant TBD
3. Document Processing Pipeline and Data Integration – Mingrui Gao (Intern)
4. Frontend – Anton Stepanenkov (Contract) and Rashan Smith (UNV)
5. GIS Processing Module – Daniel Fassbinder (Graduate researcher)
6. Knowledge Graph – Amna Mubashar (Graduate researcher)
7. Document Compilation and Cleaning – UNVs (Zijing Wang)

## Next Steps

1. **Finalize Project Scope and Objectives:** Define the specific functionalities, features, and goals of the Energy AI and Knowledge Graph Integration project in detail.
2. **Detailed Requirement Analysis:** Conduct a comprehensive analysis of user requirements, ensuring a deep understanding of the needs and expectations of end-users and stakeholders.
3. **Data Collection and Preparation:** Begin collecting data from various sources identified in the research phase. Clean, preprocess, and organize the data for integration into the knowledge graph.
4. **Knowledge Graph Development:** Start developing the knowledge graph, including defining entities, relationships, and schema. Implement algorithms for data extraction, mapping, and validation.
5. **User Interface Design and Development:** Design the user interface for querying the integrated system. Develop frontend components, incorporating query processing, data visualization, and document viewing features.
6. **Testing and Quality Assurance:** Conduct rigorous testing of the entire system, focusing on accuracy, responsiveness, user experience, and security. Address any bugs or issues identified during testing.
7. **Deployment and Monitoring:** Prepare the infrastructure for deployment, ensuring servers, databases, and security measures are in place. Deploy the system to a secure production environment and implement real-time monitoring for performance optimization.
8. **User Training and Feedback:** Provide training sessions to end-users and stakeholders for effective utilization of the system. Gather feedback and conduct usability testing to identify any areas for improvement.
9. **Continuous Optimization:** Continuously monitor the system's performance, user interactions, and feedback. Implement optimizations and improvements based on real-time data and user suggestions.

# Appendix

1. Example of entities and relations for the Knowledge Graph
2. **Entities**
3. **Energy Sources:** Detailed categories including Fossil Fuels, Renewable Energy sources (Solar, Wind, Hydroelectric, Geothermal, Biomass).
4. **Energy Technologies:** Specific technologies like Photovoltaic Cells, Wind Turbines, Hydroelectric Dams, Geothermal Systems, and Bioenergy Plants.
5. **Energy Storage:** Covering Batteries, Pumped Storage, and Thermal Energy Storage methods.
6. **Energy Consumption Sectors:** Encompassing Residential, Commercial, Industrial, and Transportation sectors.
7. **Energy Policies:** Including governmental regulations, incentives, and standards.
8. **Environmental Impact:** Quantifying greenhouse gas emissions, air, and water pollution levels associated with different energy production methods.
9. **Relations**
10. **Energy Policy and Development Environment Connections:**
    * *Energy Legislation → Renewable Energy Incentives:* Illustrates how specific energy policies incentivize the adoption of renewable energy sources.
    * *Development Agencies → Sustainable Energy Projects:* Details global sustainable energy initiatives fostering technology transfer and capacity building.
    * *Sustainable Development Goals (SDGs) → Energy Initiatives:* Establishes connections between SDGs and energy initiatives, highlighting the importance of sustainable energy solutions in achieving these goals.
11. **UN System User Group Connections:**
    * *UNDP → Energy Access Programs:* Describes UNDP initiatives providing energy access to underserved communities, promoting social and economic development.
    * *International Energy Agency (IEA) → Energy Policy Recommendations:* Highlights IEA's influence on national energy policies and strategies.
    * *World Bank → Renewable Energy Investments:* Focuses on World Bank investments supporting countries in transitioning to sustainable energy sources and technologies.
12. **Energy Policy and SDG Interconnections:**
    * *Energy Legislation → SDG 13 (Climate Action):* Demonstrates how energy policies contribute to mitigating climate change.
    * *Renewable Energy Incentives → SDG 7 (Affordable and Clean Energy):* Shows how incentives promoting clean energy align with SDG 7’s goals.
    * *Energy Efficiency Programs → SDG 9 (Industry, Innovation, and Infrastructure):* Illustrates how energy-efficient practices align with SDG 9’s focus.
13. **Platform Architecture and Technical Details:**

**1. Data Pipeline:**

The Data Pipeline is a critical component of the Energy Moonshot AI Platform responsible for acquiring and processing essential energy-related documents. It starts with the "Document Compilation - UNDP SEH," a repository containing document metadata and download links. The pipeline automates the process by downloading the documents, extracting the text, and processing the data into a structured format.

* GitHub Repository: <https://github.com/MingruiGao666/PDF-data-pipeline>
* Workflow:
  + Input: The starting point is the ["Document Compilation - UNDP SEH,"](https://docs.google.com/spreadsheets/d/1qxb6JL9f-UxLmj8dWVrWa0H8Lx4Y_HMbew35w4W25N4/edit#gid=2111530845) which includes information about various documents and links to download them.
  + Processing Steps:
    - Downloading the files from the provided links.
    - Extracting text from PDF files using the pypdfium2 library.
    - Cleaning and processing the extracted text using NLTK
      * noise removal
      * punctuation removal
      * converting text to lowercase
      * lemmatization
  + Output Format: The processed data is structured into a format that includes details such as the filename, index, country name, country code, category, document title, existence status, publication date, start year, end year, language, and content.

**2. Backend and Frontend:**

The Energy Moonshot AI Platform consists of both backend and frontend components, working together to provide a seamless user experience. The frontend, built with Next.js, is the user-facing part of the platform. It allows users to interact with the platform via a user-friendly interface. The backend, powered by Flask (a Python web framework), serves as the core engine of the platform. It offers a set of endpoints for user interactions,

* Front-End (Next.js):
  + GitHub Repository: <https://github.com/UNDP-Data/fe-energy-ai>
  + Key Features: This frontend component is developed using Next.js and provides the user interface for interacting with the Energy Moonshot AI Platform.
* Backend - Flask App (Python):
  + GitHub Repository: <https://github.com/UNDP-Data/dsc-energy-ai-backend>
  + Key Features: The backend, built using Flask and integrated with PandasAI to generate responses for the user prompt. The endpoint takes table\_name and prompt as parameters, returns text responses, table, or figures in JSON format.

**3. Semantic Search Model:**

The Semantic Search Model is a critical component of the Energy Moonshot AI Platform, enhancing its capability to understand and respond effectively to user queries. It involves user query analysis, country entity identification, text embeddings utilization, similarity search with Faiss, and generative responses using fine-tuned GPT3.5, all working together to streamline interactions and provide users with relevant information and insights.

* GitHub Repository: <https://github.com/MingruiGao666/PDF-data-pipeline>
* User Query Analysis:
  + Filter relevant documentation information.
  + Country Entity Identification (Spacy).
* Embeddings:
  + OpenAI's 'sdgi-embedding-ada-002' for text embeddings.
  + Faiss for similarity search.
* Generative Response:
  + Used the GPT3.5 model ('sdgi-gpt-35-turbo-16k').
  + Fine-tuned to generate informative responses.

1. Budget for Expanded Scope

If additional funds are available, a more robust approach can be taken for each part of the development including working with more expert consultants for each step. This expanded scope would include:

|  |  |  |
| --- | --- | --- |
| Title | Budget | Expanded Scope |
| NLP Consultants | $50,000 | There is significant potential for expanded scope throughout each of the processing modules to add intelligence and expanded case handling. The user query processing approach and content synthesis engine are valuable priorities. |
| Project Manager | $20,000 | This role will expand with the expanded team to coordinate workplans, lead communications, and review deliverables |
| Data Scientist | $10,000 | The data science module could be significantly enhanced to provide more capabilities for handling data related queries and integrating more data streams |
| Designer | $4,000 | The interfaces for interacting with documents, data, responses, queries suggestions, etc. can be improved with a designer and graphic artist |
| Expert Reviewers for the Knowledge Graph | $50,000 | **The quantity of information possible to include in the knowledge graph will be proportional to the time invested by expert reviewers to validate this information** |
| Web Developers | $12,000 | Functionality of the interface could be scaled up with expanded focus on data visualizations, user interactivity, mobile tools, and backend engineering |
| GIS Data Engineer | $20,000 | An expanded focus on the GIS integration Process GIS data, develop GIS modules, integrate with frontend. |
| Document Compilation Manager | $10,000 | The document manager will curate the document compilation database and processing pipeline, including quality assurance of all text, and curation of query examples. |

|  |  |  |
| --- | --- | --- |
| Total: | $184,000 |  |