THEME : **Mystery Theme**	
Title	*Autonomous Document Intelligence*
The Problem	Enterprises possess vast repositories of unstructured and semi-structured documents holding immense potential value, but transforming this raw information into structured, analysis-ready datasets remains a major bottleneck. Traditional IDP systems are often too rigid, focusing on predefined extraction tasks for individual documents rather than synthesizing information across large corpora to build tailored datasets. Manually curating datasets is slow, costly, and unscalable. Furthermore, the specific requirements for a dataset often evolve, or different users need differently structured views of the same underlying information. The core challenge is creating an intelligent system that not only processes documents but **allows users to dynamically define, refine, and generate specific datasets from vast document collections using simple natural language commands**, moving beyond static processing to interactive, on-demand data curation.
Expected Solution	Develop an Intelligent Document Processing (IDP) solution powered by Agentic AI, designed to **autonomously create structured datasets guided by user interaction via natural language**. The solution should follow these steps: 1. **Natural Language Driven Curation Framework:** Implement the system using an agentic AI framework (e.g., LangChain/LangGraph, AutoGen, CrewAI). Crucially, this framework must include agents capable of interpreting user requests expressed in **natural language** to define the goals, scope, schema, and filtering criteria for the desired dataset. 2. **Deep Understanding for Flexible Structuring:** Employ agents with advanced multimodal and reasoning capabilities to deeply understand content and relationships *within* and *across* documents, specifically identifying information relevant to the *current user query*. 3. **Query-Adaptive Extraction & Synthesis:** Agents dynamically adjust their extraction and synthesis strategies based on the natural language query. They extract necessary data points,

	resolve conflicts, and structure the information according to the target dataset parameters *derived from the user's request*. 4. **On-Demand Dataset Assembly & Quality Control:** Design agents to assemble the curated data points into a cohesive dataset conforming to the user-specified requirements. These agents perform automated quality checks relevant to the query (consistency, completeness, accuracy) and handle exceptions. 5. **Dynamic Schema Adaptation via Conversation:** Agents interact with the user (potentially through conversational interfaces) to clarify ambiguities in the natural language query, suggest schema refinements, and adapt the dataset structure iteratively based on feedback, allowing users to easily **change how they want the dataset structured or what data it includes using natural language.** 6. **Continuous Learning & Contextual Refinement:** Integrate feedback mechanisms allowing agents to learn from user interactions, corrections on generated datasets, and new documents, improving their ability to interpret queries and generate relevant, high-quality datasets over time. 7. **Responsible AI Integration:** Incorporate deidentification agents, guided by user specifications or policies, to create privacy-preserving datasets when requested via natural language, ensuring responsible data handling.
Resources	• Agentic AI Frameworks (e.g., LangChain, LangGraph, AutoGen, Semantic Kernel, CrewAI) • Large Language Models (LLMs) – especially those strong in instruction following and NLU. • Advanced OCR Technologies • Natural Language Processing (NLP) & Understanding (NLU) Libraries
Abbreviations	IDP: Intelligent Document Processing OCR: Optical Character Recognition NLP: Natural Language Processing NLU: Natural Language Understanding LLM: Large Language Model ML: Machine Learning API: Application Programming Interface