Paper ML02

Development of USDM through translation of human-readable protocols

Jasmine Kestemont, argenx, Ghent, Belgium
Stijn Rogiers, argenx, Ghent, Belgium
Sandeep Juneja, argenx, Cary, USA

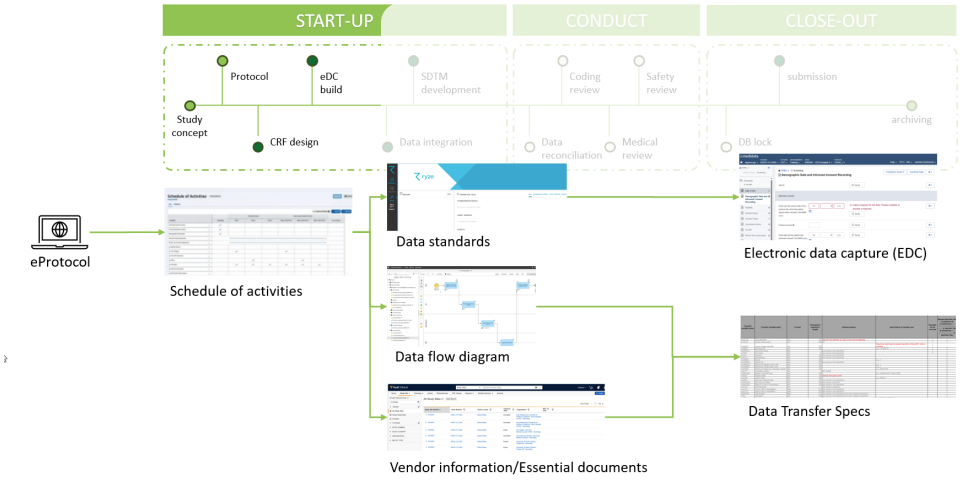
ABSTRACT

With the introduction of the Unified Studies Definition Model (USDM), there is scope to translate diverse human-readable study protocols into standardized format and develop a library that can be used to write future protocols. The information can be extracted using Natural Language Process (NLP) Language Interpretation for textural Information (LITI) rules. In order to ensure LITI rules can work for variation in protocols text, fair number of rules need to be developed, a time-consuming and complex process. Similarly information can be extracted also using Large Language Models (LLM). Developing an end-to-end process to extract information can be complicated and LLMs have risk of hallucination. However, using combination of NLP and LLMs can reduce number of LITI rules and can also help validate LLM response their by reducing hallucination risk. *In this paper we talk about how we used combination of Natural Language Processing (NLP) and Large Language Models (LLM) to identify, extract, translate and populate information from PDF protocol documents into the USDM data model.*

OUR AMBITION

FROM PROTOCOL AUTHORING TO SUBMISSION

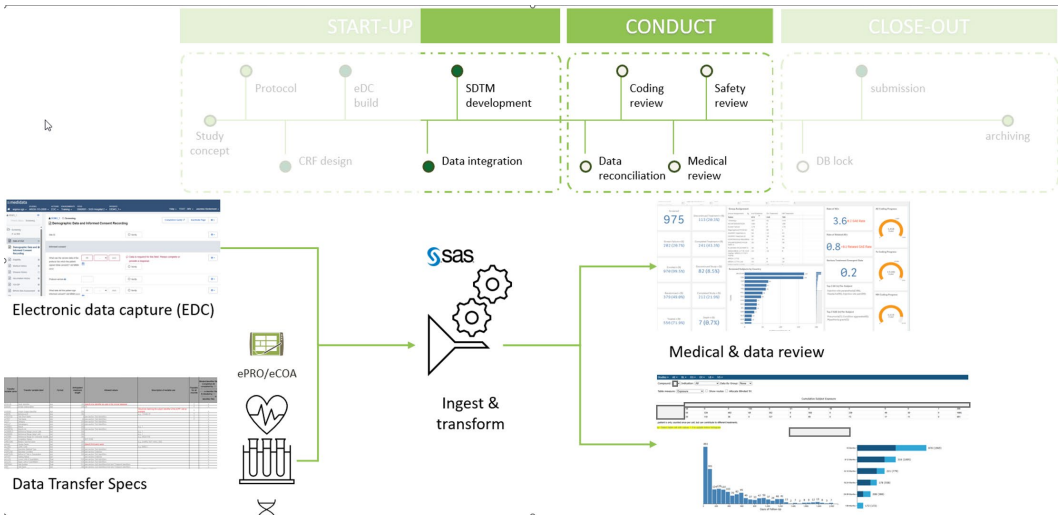
The implementation of an electronic protocol (eProtocol) in clinical trials is set to revolutionize the entire process by enabling end-to-end automation, from study design to submission to regulatory authorities. By digitizing the clinical trial protocol, eProtocols allow seamless integration across different stages of the trial, including data collection, analysis, and reporting. This automation minimizes human error, accelerates timelines, and ensures real-time data accuracy, which are critical for regulatory submissions. Moreover, it facilitates standardized workflows, enhances traceability, and improves collaboration between stakeholders, ultimately speeding up the approval process. The shift to electronic protocols will streamline compliance with regulatory requirements, making the submission process faster, more transparent, and less prone to delays.



In the near future, the eProtocol provides the framework for the schedule of activities (SoA). Automated tools help convert the study concept (sheet) into a formal protocol and guide the design; accelerate the build of the case report forms (CRFs) using Data Standards. Data flow diagrams are established digitally, ensuring consistency and alignment with regulatory requirements. The eProtocol also allows for the automatic generation of key trial documents and Data Transfer Specifications.

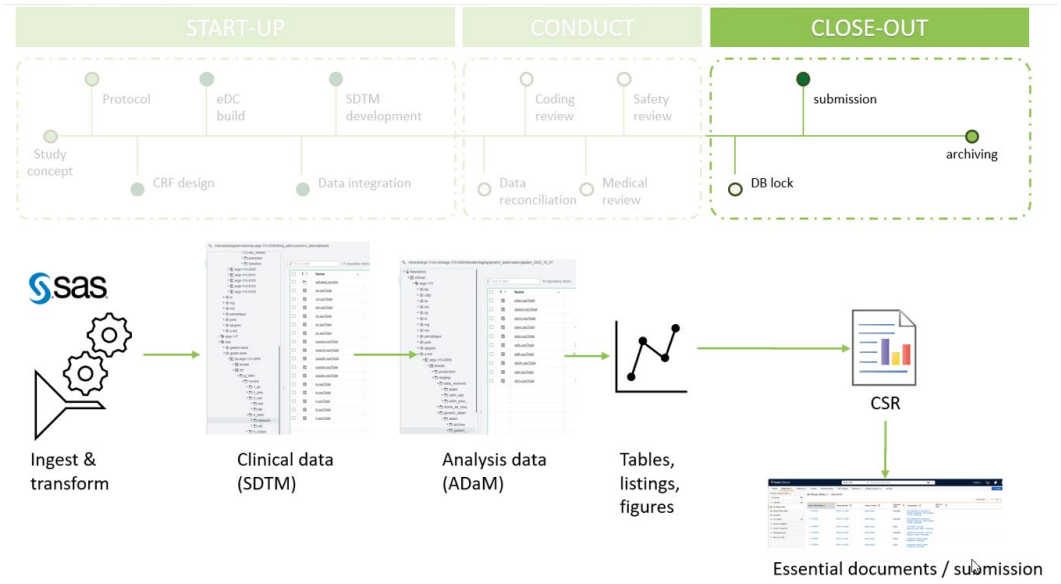
Today, we do have already at argenx the next step of the end-2-end automation in place with the successful implementation of our Near Real Time Data flow (2023-2024). EDC/eCRF data is pulled in nightly; external data (Central Lab, ECG, eCOA, etc.) weekly in our GxP qualified SAS Life Science Analytics Framework (SAS LSAF) environment.

Validated workflows trigger the automated conversion of the raw data into generic SDTM (gSDTM) nightly. These gSDTM data are being loaded into our Data Visualization & Collaboration Platform (eClinicalSolutions – elluminate). Communication between internal and external stakeholders can be done within the same elluminate platform to replace many trial oversight (XLS) trackers and be inspection ready at all times.



From these gSDTM data our Quantitative Sciences (QS/Stats Programming) colleagues generate via automated way in the same SAS LSAF platform the generic ADaM (gADaM) data that is being used for the internal Insights Data Visualization tool and TLFs. For more detailed information see “PHUSE EU Connect 2023 – DV09 – Oversight and Insight into a fast moving Biotech company”.

Finally, digital protocols can transform how clinical trial data is consolidated and reported in the final Clinical Study Report (CSR). Traditionally, the preparation of a CSR involves manually pulling data from various sources, organizing it into a coherent narrative, and ensuring compliance with regulatory requirements. Digital protocols simplify this process by offering several key advantages: **Automated Data Integration; Template-Based Report Generation and AI-Assisted Writing and Analysis.**



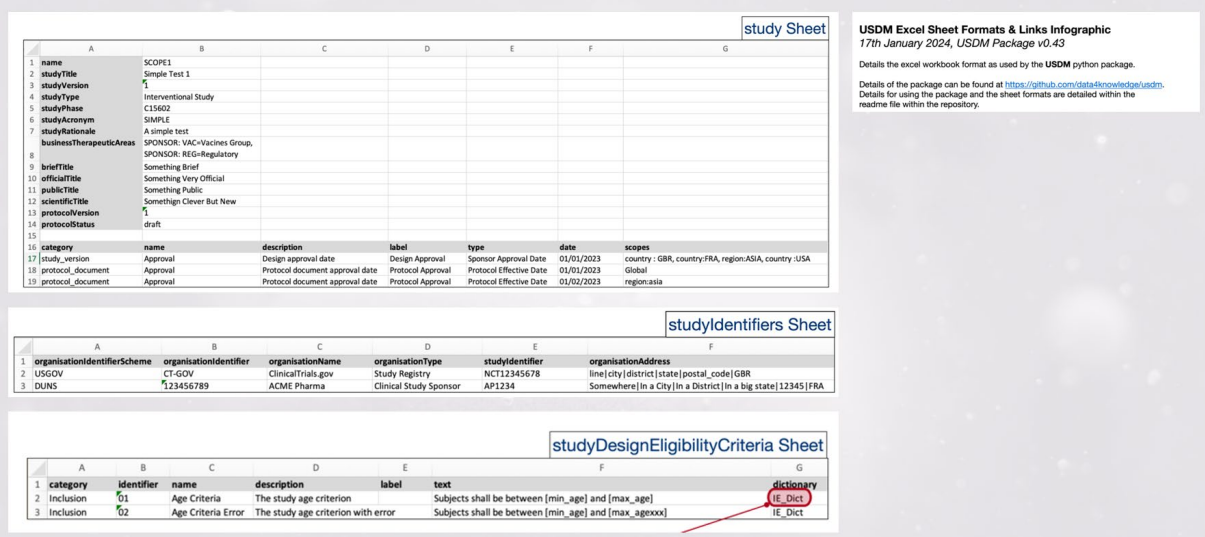
POC METHODOLOGY

PROTOCOL TRANSLATION

Employing NLP & LLM algorithms, we developed a Proof of Concept (POC) to automatically translate human-readable study protocols into a structured format. This involves extracting key information such as study objectives, methodologies, inclusion/exclusion criteria, and outcomes.

MS OFFICE INTEGRATION

Excel's ubiquity and user-friendly interface makes it an ideal platform for implementing the Unified Studies Definition Model (USDM). The translated protocols are integrated into Excel templates, providing a familiar environment for researchers to input and manipulate data. The templates adhere to a standardized structure defined by the USDM, ensuring consistency across diverse research domains.



PHUSE EU Connect 2023 - DDF Workshop - PUBLIC - Wiki (cdisc.org)

EXPECTED OUTCOMES USDM

- ✓ **Improved Collaboration**
  The Unified Studies Definition Model will promote collaboration by providing a standardized framework for communicating and sharing study protocols, thus fostering a more efficient and transparent research environment.

- ✓ **Enhanced Data Interoperability**
  The standardized structure of the USDM will enable seamless integration and comparison of data across studies, promoting meta-analyses and systematic reviews.

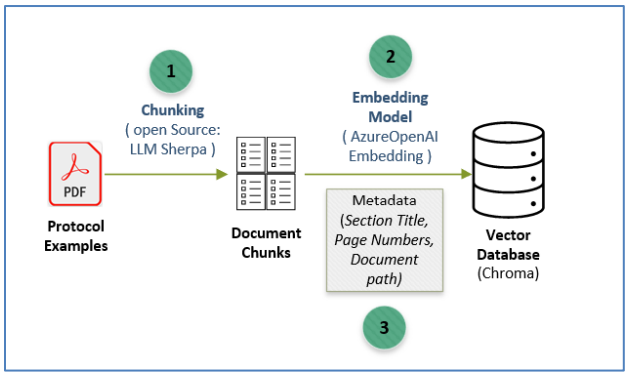
- ✓ **Time and Resource Efficiency**
  By automating the translation of protocols, researchers will save time and resources that would otherwise be spent on manual data extraction and interpretation.

- ✓ **Increased Reproducibility**
  The transparency and standardization introduced by the USDM will enhance the reproducibility of studies, contributing to the overall reliability of scientific research.

POC APPROACH

The process was broken down into two phases:

PHASE 1: PRE-PROCESSING DOCUMENTS

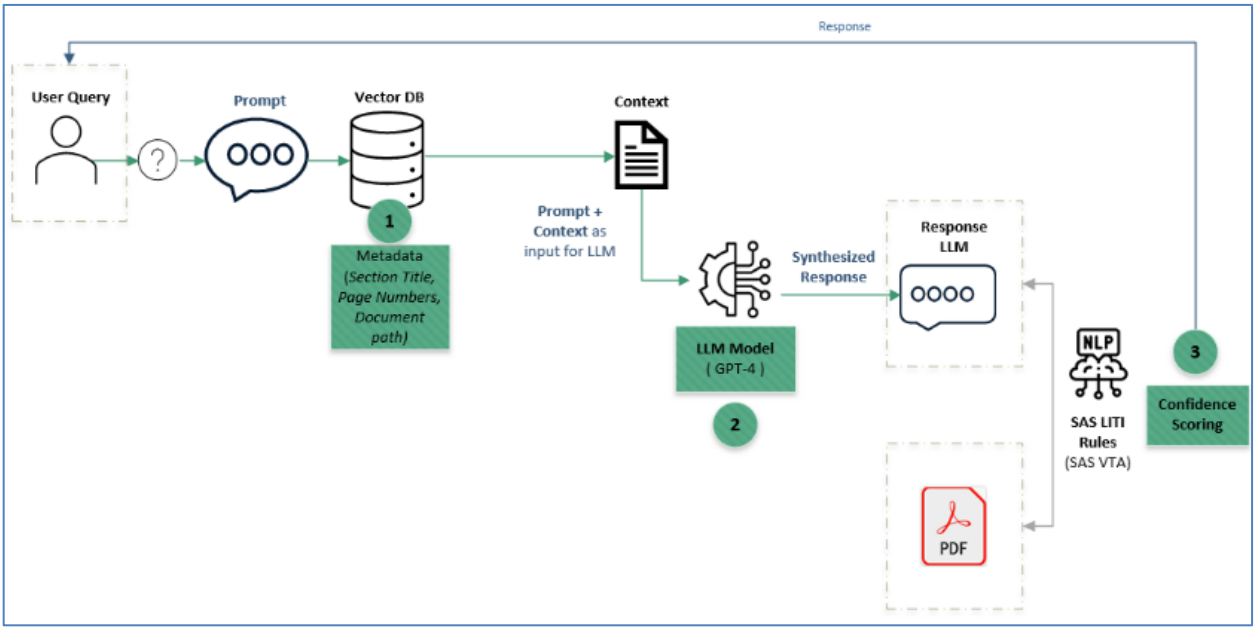


1.1 Chunking (open source - LLM Sherpa):
The process begins by parsing protocol PDF documents and breaking them into smaller, manageable chunks of information.

1.2 Embedding Model (Azure OpenAI Embedding):
The document chunks are then processed by an embedding model, which converts the text into embeddings (numerical representations) and associates metadata like section titles, page numbers, and document paths.

1.3 Vector Database (Chroma): The embeddings and their associated metadata are stored in a vector database, which enables fast and efficient similarity searches. This allows the system to quickly retrieve relevant document chunks in response to queries based on semantic similarity.

PHASE 2: ANSWER GENERATION WITH ACCURACY ASSESSMENT



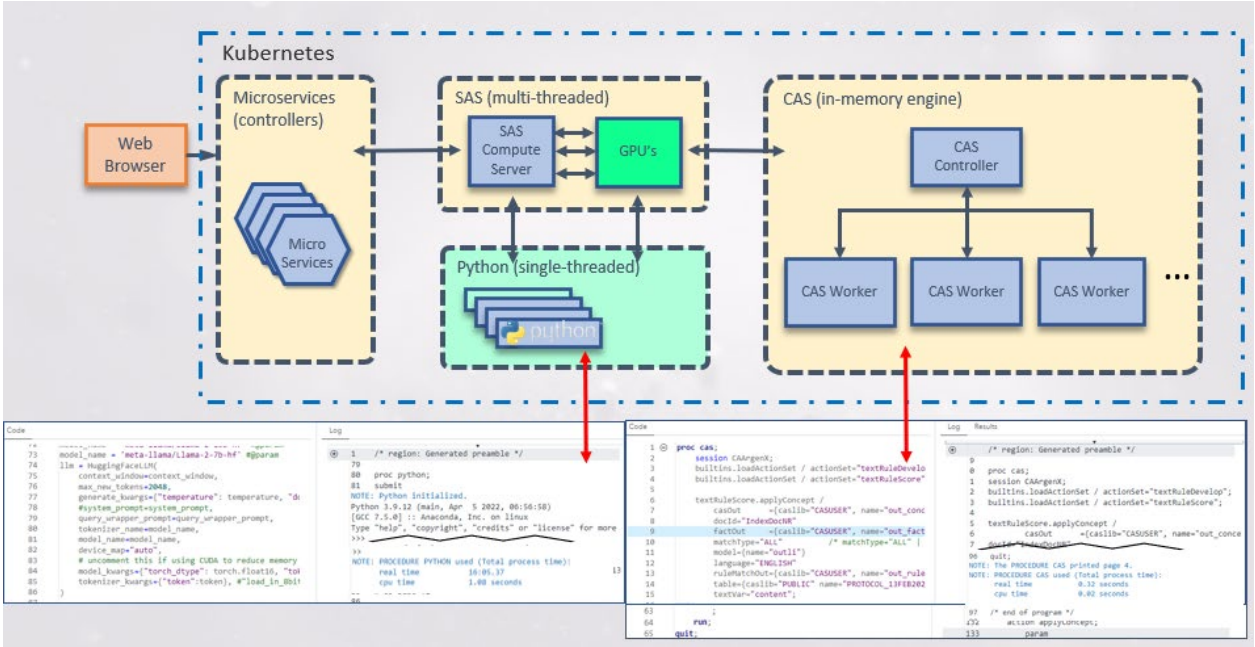
2.1 User Query and Metadata Retrieval (Vector DB):
The user inputs a query, and the system searches the Vector Database, which contains document chunks along with their metadata (section titles, page numbers, document paths). This retrieves relevant context for the user's query.

2.2 LLM Model (GPT-4):
The prompt from the user query and retrieved context is fed into a Large Language Model (GPT-4), which synthesizes a response based on the context provided. This ensures that the answer is both accurate and contextually relevant.

2.3 Confidence Scoring (SAS LITI Rules):
After the response is generated, natural language processing (NLP) rules (i.e., SAS LITI) are applied to assess the confidence level of the response. This final score helps determine the reliability of the generated answer.
(https://blogs.sas.com/content/subconsciousmusings/2024/05/01/five-steps-to-improve-information-extraction-using-trustworthy-generative-ai/)

THE ENVIRONMENT (SAS VIYA, OPEN SOURCE AND OPEN AI)

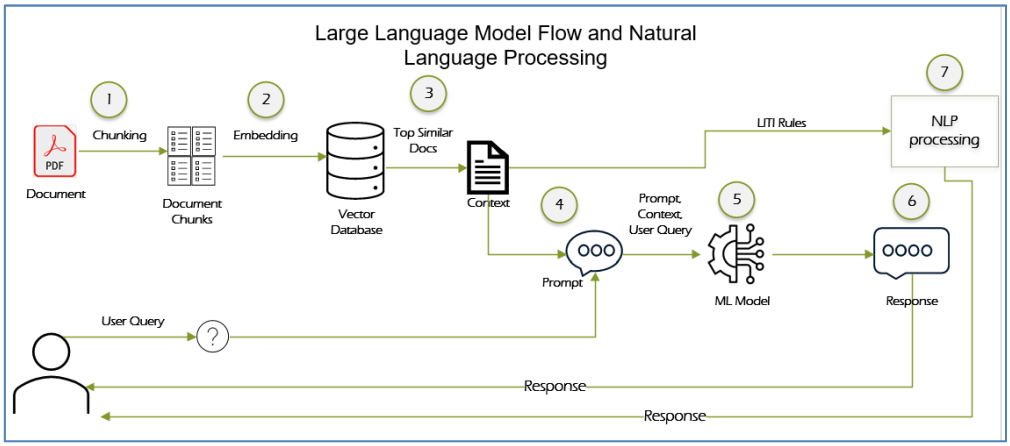
SAS Viya, integrates with open-source tools and OpenAI, enabling powerful, scalable analytics in a flexible environment. This combination fosters innovation while maintaining the reliability and governance of enterprise-level solutions.



CHALLENGES IN LLM & NLP PROCESS

1. Data Pre-processing & Chunking: Extracting metadata, page summaries, and processing tables/images from complex, unstructured documents can be inconsistent and error-prone, requiring advanced techniques for accurate parsing.

2. Data Embedding: Embedding models like BERT, GPT, or custom embeddings capture semantic meaning, but finding the most suitable model for specific applications is difficult. Selecting and fine-tuning the right embedding model for semantic understanding is challenging, as different models vary in performance based on the task and domain.



3. Top Similar Documents Retrieval: Ensuring efficient vector search over large datasets requires balancing between precision and computational cost, as well as optimizing the distance metric to capture semantic similarity accurately.

4. Prompt Engineering: Designing effective prompts and tuning parameters to guide LLMs is often a trial-and-error process, prone to inconsistencies and model biases.

5. ML Models (LlaMa-7b, OpenAI): Smaller models like LlaMa-7b may struggle with complex tasks, while larger models like OpenAI can be resource-intensive and prone to hallucinations.

6. Synthesizing Response: LLMs may struggle with coherently synthesizing responses from multiple sources. Challenges include ensuring consistency, handling contradictions, and maintaining factual accuracy, especially when merging information from diverse documents.

7. NLP LITI Rules: Writing generalizable rules for NLP tasks across different contexts and documents is difficult due to the complexity and variability of natural language.

NEXT STEPS - THE JOURNEY CONTINUES ?

SHORT TERM (Q2-Q4 2024)

Leverage OpenAI Service [completed]
Utilize commercial OpenAI models instead of LLAMA-V2 7B while finalizing the internal business case.

Reflect on POC [ongoing]
Review proof of concept (POC) results and establish clear next steps by end Q4 2024.

MID-TERM (2025)

Focus on CTPS
Prioritize the development of the Concept Sheet (CTPS) for strategic alignment.

Schedule of Assessments (SoA):
Implement the Schedule of Assessments to monitor progress and key milestones.

LONGER TERM – DIGITAL PROTOCOL

Cost: Carefully assess the financial investments vs. impact.

Technology: We are biotech, not software builders. Off-the shelf solutions available / Co-creation ?

Change Management: Implement strategies for managing organizational changes.

New Roles: Define and introduce new roles to support the implementation and usage of Digital Protocols.

ACKNOWLEDGEMENTS
We would like to thank Petra Van Vlierberghe (VP Clinical Development, argenx) for the continuous support and leadership.

CONTACT INFORMATION
Your comments and questions are valued and encouraged.

Contact the authors at:

Jasmine Kestemont
jasmine.kestemont@innovion.be
Jasmine Kestemont | LinkedIn

Stijn Rogiers
srogiers@argenx.com
Stijn Rogiers | LinkedIn

Sandeep Juneja
sjuneja@argenx.com
Sandeep Juneja | LinkedIn