AI Curation Methods for NASA Scientific Data

Walter Alvarado[1], Lauren Sanders[1], Hari Parthasarathy[2], Harlan Phillips[2], Kayvon Crenshaw[3], Samrawit Gebre[1], Sigrid Reinsch[1], James Casaletto[4], and Sylvain Costes[1]

[1]NASA Ames Research Center, Mountain View, CA, [2]University of California, Berkeley, Berkeley, CA, [3]University of Illinois at Urbana-Champaign, Champaign, IL, [4]Blue Marble Space Institute of Science, Seattle, WA

Abstract

The AI for Curation project aims to integrate advanced LLM models into various aspects of our data curation workflow, enhancing the efficiency and accuracy from submission to user interaction. This initiative will impact multiple touchpoints, including data ingestion, curation processes, and user engagement with our curated datasets, affecting various centers, domains, and a broad user base. First, we are developing tools capable of parsing incoming data across all formats and utilizing LLMs to convert unstructured data into structured, standardized formats. This process streamlines curation by converting data into community-standard formats like ISA-Tab, significantly reducing the time and effort required by curators. This allows curators to allocate more resources to scientific analysis rather than data formatting tasks. Second, we are implementing AI/ML models to automate and enhance the accuracy of data validation and verification. These models ensure that the curated data adheres to high standards of quality and reliability, benefiting researchers and data users by providing them with rigorously verified datasets. Finally, we are developing a conversational agent (chatbot) that interfaces with our extensive repository of curated scientific studies on the Open Science Data Repository (OSDR). The chatbot enhances data discoverability by assisting users in navigating the knowledge base and referencing relevant studies. This improvement in accessibility makes scientific data more available to the community, thereby promoting the principles of open science.

Structured Outputs from LLM

In our efforts to enhance the AI for Curation project, we've developed advanced tools for automated tagging that significantly streamline the curation process. These tools are adept at identifying patterns within diverse datasets, automatically suggesting context-relevant metadata tags. This capability not only facilitates the initial organization of newly ingested data but also allows for the retrospective curation and categorization of existing datasets. Once data is aligned with community standards, these tools apply contextual tags to meticulously organize the information, rendering it AI-ready for subsequent training processes. For example, they can generate specific tags for studies involving astronauts or particular experimental conditions, ensuring a high level of precision and relevance in data handling. This method greatly enhances the accessibility and utility of the curated data, supporting more efficient research and analysis.

```python
class Study(BaseModel):
    Title: str = Field(description="Study title.")
    Accession: str = Field(description="Study accession number")
    Cells: str = Field(description="Metadata values for 'cell line'
                                    for the study's samples, if cell line is not include
                                    in the metadata, nan is used.")
    Astronaut: bool = Field(description="True if the study was done on astronauts,
                                    False if the study was not done on
                                    astronauts.")
```

Pydantic, a Python library for data validation and settings management, facilitates the creation of schemas that ensures responses from large language models (LLMs) conform to a specified structure. Here, a Pydantic class is used to define the structure for automated tagging of OSDR studies, illustrating how data is organized and standardized for AI processing.

Data Alignment & Verification

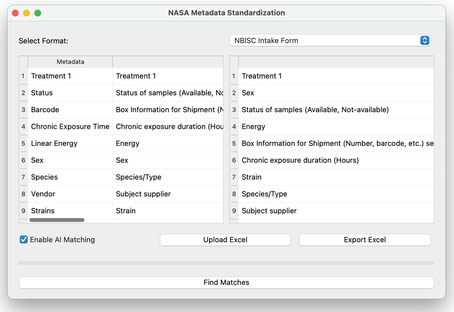
Generative AI can be applied throughout the entire curation pipeline, offering numerous use cases, especially in data curation. One of the initial challenges addressed was the alignment of data received from researchers. Data often arrives in various formats and with disparate terminologies. To tackle this, we have developed tools that map and align these terms to standardized ontologies. For instance, while fields like "sex" and "genotype" may be consistent, others may use different terms for the same concepts.

Combining programmatic techniques with generative AI provides a robust toolkit for curating datasets. For example, we utilize similarity metrics such as the Levenshtein distance to match strings that are similar, handling a significant portion of the workload. For terms that use different words but are semantically similar, we leverage generative AI to understand the context and find the appropriate matches, enhancing the precision and utility of our curated datasets.

To further streamline the data alignment process, our graphical user interface (GUI) provides a suite of intuitive tools. These GUI capabilities are integral to ensuring that the data not only meets standardization requirements but also aligns perfectly with the users' specific research needs and contexts.

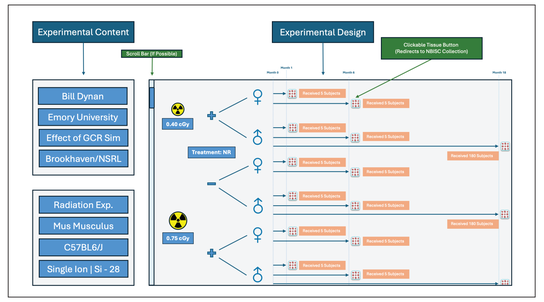
Column in File1	Best Match in File2	Similarity Score	Match Status
Euthanasia method	Euthanasia method	100	Above Threshold
Sample preservation method	Sample preservation method	100	Above Threshold
Biospecimen category	Biospecimen category	100	Above Threshold
Sex	Sex	100	Above Threshold
Genotype	Genotype	100	Above Threshold
Absorbed dose rate	Absorbed dose rate	97	Above Threshold
Number of fractions	Number of fractions	97	Above Threshold
Time point of sacrifice post irradiation	****Time point of sacrifice post-irradiation	93	Above Threshold
Additional notes	Additional Notes	91	Above Threshold
Absorbed dose rate (Unit)	Absorbed dose rate unit	88	Above Threshold
Total absorbed dose (Unit)	Total absorbed dose units	88	Above Threshold
Time between exposures (Unit)	Time between exposures (Hour)	86	Above Threshold
Absorbed dose per fraction	Total absorbed dose per fraction	86	Above Threshold
Time between fractions (Days)	Time between fractions value	84	Above Threshold
Sample storage temperature	Sample storage temperature (Celsius)	84	Above Threshold
Energy (Unit)	Energy unit	83	Above Threshold
Absorbed dose per fraction (Unit)	Total absorbed dose per fraction units	82	Above Threshold
Dose 1 (non-radiation)	No Match Found	0	Below Threshold

```
Matching: Dose 1 (non-radiation)
LLM Match: Treatment 1 dose
Match found!
 Dose 1 (non-radiation) <---> Treatment 1 dose
_____________
Matching: Status (stts_name)
LLM Match: Status of samples (Available, Not-available)
Match found!
 Status (stts_name) <---> Status of samples (Available, Not-available)
_____________
```



Users can choose from predefined templates, which promote consistency across different datasets. Once a template is selected, the system automatically identifies and matches the fields according to the template structure. This automated matching significantly reduces the time and effort required to align data.

Auto-Generated Graphical Abstracts



To enhance the presentation and understanding of experimental data, we levearge generative AI to create graphical abstracts. Users can manually input experimental parameters using a user-friendly interface, ensuring direct control over the data visualization process. Additionally, the system supports file imports, as well as intelligent extraction techniques using generative AI to parse and interpret data directly from scientific papers. This capability automates the visualization of complex experiments, transforming dense textual information into clear, informative graphical abstracts.

Enhancing Study Discoverability

We've implemented a chatbot to improve how researchers and users interact with our data. The chatbot integrates directly with our databases, pulling information in real-time to ensure that responses are always up-to-date and accurate. This facilitates dynamic interaction based on the latest data available. It also employs natural language processing to summarize intricate research findings or explain complex scientific concepts, making the information more accessible and easier to understand for a broader audience.

Additionally, the chatbot incorporates a Retrieval-Augmented Generation (RAG) model, which enhances its ability to generate precise and contextually relevant responses. The RAG model first retrieves relevant information from a vast database of scientific literature before producing a coherent reply, thereby improving the chatbot's effectiveness in navigating and discovering scientific studies. These integrated features ensure that critical information is not only more discoverable but also more understandable and secure within the scientific community.

This chatbot initiative aligns with our commitment to open science, facilitating broader access to knowledge and fostering a collaborative environment for innovation.

References

1. Colvin, S., Jolibois, E., Ramezani, H., Garcia Badaracco, A., Dorsey, T., Montague, D., Matveenko, S., Trylesinski, M., Runkle, S., Hewitt, D., & Hall, A. (2024). Pydantic (Version 2.9.0) https://github.com/pydantic/pydantic
2. Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997 (2023).
3. Dubey, Abhimanyu, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).
4. Data are courtesy of the NASA Open Science Data Repository (OSDR). NASA Ames Research Center. (2024).

We extend our thanks to the Ames Life Sciences Data Archive (ALSDA), NASA Biological Institutional Scientific Collection (NBISC), and NASA GeneLab, integral projects of the Open Science Data Repository (OSDR), for their invaluable contributions to this project. Additionally, we are grateful to the AI for Life Science (AI4LS) group for their innovative approaches that have significantly enhanced our data analysis processes. Our appreciation also goes out to the Analysis Working Group (AWG), a team of dedicated volunteers, whose expertise and commitment have been instrumental in refining our research outputs.

NASA SILICON VALLEY AMES RESEARCH CENTER