Leveraging AI Agents for Efficient Data Annotation

Chang In Moon[1], Anh Nguyet Vu[1], Alex Verbitsky[2], Brian Dennis[2], Chris Harman[2], Mark Weston[2], Mohammed Eslami[2], Patrick Boutet[2], Christina Parry[1,*], Jineta Banerjee[1], John Hill[1], Jay Hodgson[1], Alberto Pepe[1], Milen Nikolov[1], Sonia Carlson[1] and Robert J. Allaway[1]

[1]*Sage Bionetworks, Seattle, WA, USA*
[2]*Neritas, Annapolis, MD, USA*

Abstract

Effective management and analysis of biomedical datasets are critical components for accelerating translational research and advancing healthcare innovations. Well-structured metadata enables data to be easily searchable, understandable, and reusable, thus facilitating reproducibility and advancing scientific discovery. However, tasks related to curation and metadata annotation are labor-intensive tasks. For example, the standard process for data ingress and curation involves users uploading data to a storage platform and manually annotating each file with metadata according to a specific data model, typically through spreadsheets or a user interface. This process is burdensome particularly when biomedical datasets contain hundreds of files. This time-consuming process requiring specialized knowledge, can lead to significant delays in making data findable. Consequently, these challenges can result in missed opportunities for researchers to re-use the data. Sage Bionetworks is the developer of a platform called Synapse which hosts more than 3 petabytes of biomedical research data. In this demonstration, we present Sage Bionetwork's effort to streamline the data annotation process using artificial intelligence (AI). The demo highlights the Synapse Agent and its transformative role in metadata annotation and data integration workflows. The Synapse Agent automates metadata annotation tasks by leveraging schema-bound file entities and contextual information from project wikis, enabling rapid and accurate assignment of metadata. It performs both simple tasks, such as extracting valid values from filenames, and complex inferential annotations guided by user input or schema rules. This demonstration underscores Sage's commitment to creating scalable, FAIR-compliant data management solutions for life sciences. Attendees will gain insights into the practical implementation of AI-driven tools for metadata annotation and data harmonization, offering a blueprint for addressing challenges in data integration.

Keywords

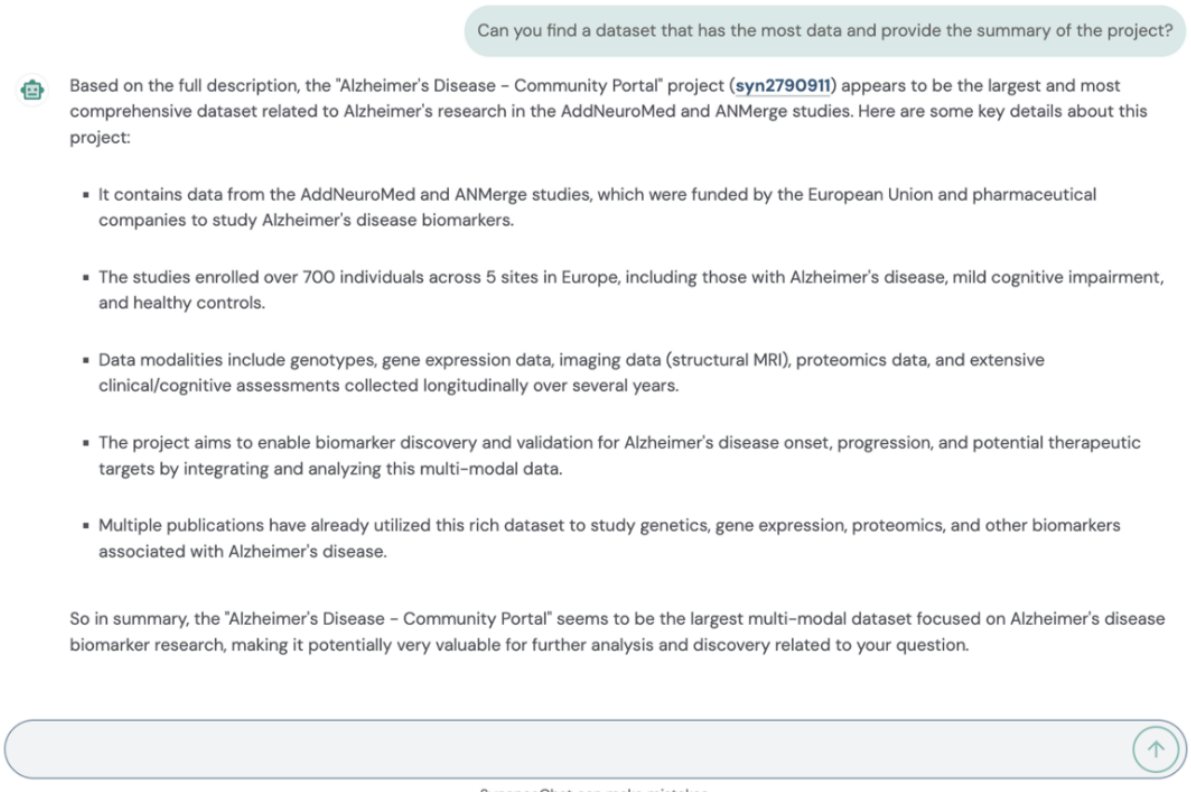
biomedical data management, metadata annotation, data curation, artificial intelligence,

1. Introduction

Biomedical research increasingly relies on large, complex, and heterogeneous datasets. Effective annotation and integration tools are critical for unlocking the potential of these datasets [1], driving translational research, and facilitating reproducible science. To address the time-consuming nature of metadata curation, Sage Bionetworks has developed innovative approaches to automate and streamline these processes. At the core of these efforts is the Synapse platform, a robust data and project management system hosting over 3 petabytes of biomedical research data. As part of our broader initiatives, we created a proof-of-concept AI application that integrates a chat-based large language model (LLM) agent to assist researchers with metadata annotation and data integration tasks. The chat agent is equipped with specialized functions that allow it to interface directly with Synapse, enabling researchers to streamline complex workflows. The agent's key capabilities include performing searches and navigation across Synapse to locate datasets, project components, and forum posts, as well as managing metadata by extracting and annotating metadata from file entities, annotation views, and table schemas. This agent enhances researchers' productivity by automating repetitive tasks and facilitating inferential reasoning for complex workflows. For instance, it can extract valid values from filenames or infer metadata annotations based on schema rules and user- provided input. Figure 1 illustrates a practical

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.

Figure 1: Synapse Chat Agent Querying Alzheimer's Disease Datasets on the Synapse Platform.



use case where the agent assists a user in exploring and retrieving datasets from projects on the Synapse platform, highlighting its ability to streamline data navigation and discovery.

We will demo several features of the upcoming Synapse Chat Agent that automate key tasks for metadata annotation, as shown in Figure 2. These tasks include editing and managing annotations on Synapse entities, extracting and translating metadata from various input formats into target schemas, and validating metadata accuracy by aligning it with predefined models. Additionally, the agent can perform sense-checking to ensure metadata consistency across project annotations and input data, as well as expanding metadata models to incorporate new attributes and concepts. This demonstration highlights the Synapse Chat Agent's ability to streamline repetitive metadata curation workflows, harmonize complex datasets, and support FAIR-compliant data management. By automating manual processes and enhancing semantic reasoning, the agent empowers researchers to focus on generating insights and driving innovation. Future enhancements will further expand its capabilities to handle diverse datasets and workflows, solidifying its role as a critical tool for accelerating precision medicine and scientific discovery.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. Hulsebos, W. Lin, S. Shankar, A. Parameswaran, It took longer than I was expecting: Why is dataset search still so hard?, in: Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics, ACM, New York, NY, USA, 2024, pp. 1–4.

Figure 2: Synapse Chat Agent AI-Auto Annotation Workflow Tasks