DGLink: automated knowledge graph construction from biomedical data repositories

Woodward Galbraith[1], Benjamin M. Gyori[1,2]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA

[2]Department of Bioengineering, College of Engineering, Northeastern University, Boston, MA 02115, USA

1 Introduction

There are numerous data repositories maintained by funding agencies and private institutions that aim to provide access to user-submitted multi-modal experimental and clinical datasets. The data in these repositories hold substantial scientific value. However, in practice, incomplete annotation and a lack of standardized formats are critical bottlenecks limiting their use at scale. The NCI General Commons (GC) [1] for example, provides access to data from 41 oncology studies spanning over 75,000 participants and 600,000 files across diverse modalities, yet faces exactly these challenges. To address this gap, we have developed DGLink, an automated system that traverses a portal's studies and experimental datasets to construct semantic annotations for experimental conditions and readouts. Existing semantic search systems such as DUG [2] aim to extract semantic annotations from study-level metadata. DGLink extends these approaches by extracting semantic annotations from both a study's metadata and its associated experimental files. This is accomplished through the schema-free ingestion of both experimental tabular files and modality-specific experimental files (e.g., VCF, DICOM formats) coupled to named entity recognition and normalization (grounding) against biomedical ontologies. These annotations as well as portal-specific structural information and study metadata are then assembled into a knowledge graph (KG), offering two main benefits: (1) The KG serves as a semantic interoperability layer across a portal's studies, increasing findability and reusability. (2) The KG connects a portal's studies to external knowledge, enabling knowledge-driven interpretation. We demonstrate DGLink on the Neurofibromatosis Data Portal [3], the NCI Genomics Data Commons [4] and NCI GC. For instance, DGLink applied to the NF Data Portal results in a KG integrating 310 disease studies yielding more than 22,500 nodes and 51,000 edges. Finally, we describe how the resulting KGs can be leveraged by a large language model (LLM) agent through a Model Context Protocol (MCP) interface for data discovery and knowledge-guided hypothesis generation. DGLink is available as an open-source Python package under the BSD 2-clause license at https://github.com/gyorilab/dglink.

2 Methods

While traversing a data portal, DGLink extracts information from two core sources: (1) Portal-specific structural information, that is, how the portal itself organizes its content and studies. (2) Project content, that is, project metadata and experimental files which are annotated using Gilda [5], which grounds entities to several key biomedical ontologies and terminologies including HGNC, UniProt, ChEBI, MeSH, DOID, HP, and EFO. During experimental file annotation, DGLink processes both tabular and modality-specific files. To accommodate the wide variance in formats across tabular experimental files, DGLink assumes no fixed file-schema, instead ingesting files with a flexible custom data-loader, and annotating each column independently using Gilda. Additionally, DGLink uses an extensible framework for extracting annotations from modality-specific file formats. Currently, only processors for DICOM and VCF file-types are implemented but this can be expanded. The extracted portal specific structural information and annotated project content are then assembled into a KG conforming to the Biolink model [6] ensuring interoperability and reproducibility. Finally the KG is instantiated in a Neo4j (https://neo4j.com) graph database.

Additionally, to reduce false positives during the annotation of a study's tabular experimental files DGLink implements mechanisms to select from each table only columns that are likely to contain relevant entities without assuming a fixed-table schema or existing annotations. DGLink implements three column selection methods: (1) A rule-based method, and (2) an LLM-based method inspired by Magneto [7] which attempts to map individual table-columns to a predefined schema of biological data types. Column selection can be done on a "per-column" basis where a separate LLM call is made for each column or on a "table-wide" basis using only one call. Additionally, the method is provider-agnostic and can be run locally using Ollama (https://ollama.com/) to accommodate non-public datasets. Finally, DGLink implements (3) a hierarchical method which runs schema matching column selection only on columns that pass the rule-based method as an initial filter.

The KG assembled by DGLink enables several query modalities for data discovery and integration with data interpretation engines: (1) directly as a Neo4j-compliant node and edge set, (2) through a structured semantic search web-interface, and (3) through an LLM-based chat interface using a Model Context Protocol (MCP) server, in which an agent draws on the KG as an interoperable representation of the portal data alongside its background scientific knowledge enabling natural language queries for data discovery and hypothesis generation.

Portal	Studies processed	Nodes	Edges	Tabular coverage
NF Data Portal	310	22,500	51,100	96%
NCI GC	2	5,500	5,800	98%
NCI GDC	1	40,000	117,500	100%

Table 1: Summary of KGs constructed by DGLink from three data portals.

Method	Accuracy	Precision	Recall	F1 score
Rule-based	0.513	0.268	0.822	0.396
Per-column schema matching	**0.920**	0.781	**0.862**	**0.819**
Table-wide schema matching	0.880	0.641	**0.862**	0.735
Hierarchical	**0.920**	**0.800**	0.827	0.813

Table 2: DGLink's column selection performance. LLM-based methods use GPT-5 mini as the base model.

3 Results

We demonstrate DGLink's high accuracy and generalizability through its application to three portals: the NF Data Portal, the NCI GC and the NCI GDC. For the NF Data Portal and NCI GC, DGLink was applied to all publicly accessible data, while for the NCI GDC a single large study was processed as a proof of concept. Table 1 summarizes the resulting KGs and shows that DGLink is able to annotate the vast majority of tabular datasets across projects (despite the lack of a shared schema), notably achieving 98% coverage of available tabular datasets on the NCI GC, which hosts data from a wide range of modalities. Importantly, despite differing modalities and APIs across portals, DGLink generalizes well, constructing semantically interoperable KGs in each case.

To benchmark column selection methods, we manually annotated 150 randomly sampled columns from the NF Data Portal as either relevant biological entities or non-entities. Table 2 shows that the rule-based method had high recall but low precision, making it a good initial filter in the hierarchical method. The two schema matching methods had higher overall metrics, with "per-column" slightly outperforming "table-wide" matching. The hierarchical method trades a small reduction in F1 for a $\approx 35\%$ decrease in LLM calls compared to per-column schema matching. Separately, manual evaluation of 100 randomly selected annotations from the NF Data Portal KG showed Gilda achieved 96% grounding accuracy.

Figure 1 shows a sub-graph of the NF Data Portal KG. This sub-graph contains portal-specific structural information (e.g. dark pink "Study" nodes), study metadata (e.g. light blue "Institution" nodes), and automatically generated annotations of a study's experimental files (e.g. navy "Drug" nodes). Critically, the drug Axitinib was only identified by traversal of experimental files and their annotation, demonstrating that file-level grounding creates semantic links between studies that would be missed by metadata-only approaches.

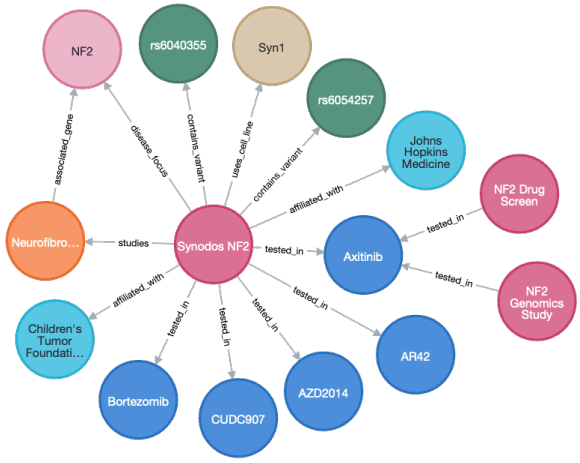

Figure 1: Sub-KG extracted from the NF Data Portal

4 Discussion

We presented DGLink, an automated system for constructing semantically grounded KGs from biomedical data portals, and demonstrated it across three portals. Key areas of future work include integrating DGLink-constructed KGs with external KGs such as ones constructed at scale from literature [8] to enable mechanistic reasoning over entities in data portals, and integrating DGLink directly with data portals for production use.

Funding statement

This work was funded by the DARPA Automating Scientific Knowledge Extraction and Modeling (ASKEM) and ARPA-H Biomedical Data Fabric (BDF) programs (HR00112220036).

References

[1] Erika Kim et al. In: *Cancer Research*. DOI: 10.1158/0008-5472.CAN-23-2730.

[2] Alexander M Waldrop et al. In: *Bioinformatics*. DOI: 10.1093/bioinformatics/btac284.

[3] Robert J. Allaway et al. In: *Scientific Data*. DOI: 10.1038/s41597-019-0317-x.

[4] Allison P. Heath et al. In: *Nature Genetics*. DOI: 10.1038/s41588-021-00791-5.

[5] Benjamin M Gyori et al. In: *Bioinformatics Advances*. DOI: 10.1093/bioadv/vbac034.
[6] Deepak R. Unni et al. In: *Clinical and Translational Science*. DOI: 10.1111/cts.13302.
[7] Yurong Liu et al. In: *Proc. VLDB Endow.* DOI: 10.14778/3742728.3742757.
[8] John A. Bachman et al. In: *Molecular Systems Biology*. DOI: 10.15252/msb.202211325.