CDEAtlas: Interactive Visualization and Exploration of Common Data Elements at Scale

Vincent J Zhang, MS[#], Ruey-Ling Weng, MS[#], Yujia Zhou, MS, Lingfei Qian, PhD, Hua Xu, PhD, Huan He, PhD[*], Na Hong, PhD[*]

Department of Biomedical Informatics & Data Science, Yale School of Medicine, New Haven, CT

Abstract

Common Data Elements (CDEs) are critical for biomedical data standardization and interoperability, yet large repositories remain difficult to explore because of their scale, heterogeneity, and limited support for understanding and interpretation. We developed CDEAtlas, an interactive platform for visualizing semantic, organizational, temporal, and reuse patterns at scale. Each CDE was embedded from its name, description, and question text and projected into a two-dimensional semantic landscape. Using CDEAtlas, we analyzed 22,743 CDEs from the NIH CDE Repository, and quantified concept coverage, steward organization contributions, CDE use patterns, temporal growth, and semantic overlaps. The results indicate limited concept standardization, uneven organizational contributions and form usage patterns, and redundancy exists across the NIH CDE Repository. CDEAtlas combines quantitative analysis and interactive visualization in one platform, allowing users to easily explore the CDE ecosystem, identify harmonization priorities, and track how CDEs evolve over time, helping improve CDE discovery, harmonization, and management.

Introduction

Despite the growing volume and importance of Common Data Elements (CDEs) in promoting data standardization and interoperability across biomedical research, navigating large-scale CDE repositories remains challenging due to their scale, heterogeneity, and limited exploratory tools[1]. Existing repositories are typically accessed through keyword-based search interfaces that support lookup of individual elements but provide limited insight into the broader CDE landscape[2]. As a result, researchers and data curators often struggle not only to identify relevant CDEs, but also to understand semantic relationships among elements, patterns of reuse across studies and organizations, and how CDE development has evolved over time[3]. These limitations create substantial barriers in CDE selection, harmonization, and reuse, particularly for interdisciplinary teams seeking overlapping or equivalent elements across different sources.

Several tools have been developed to support the search, browsing, and discovery of data elements, including the National Institutes of Health (NIH) CDE browser[4], the National Cancer Institute (NCI) CaDSR II browser[5], and ImmPort shared data[6]. Additionally, tools such as CDEMapper[7] and IMI-CDE[8] have been developed to assist researchers in mapping study variables to relevant NIH CDEs. Nevertheless, the broader discoverability challenge of understanding the CDE landscape at a macro level before examining individual elements remains largely unaddressed. Interactive visualization has proven valuable for exploring complex biomedical information, including ontologies and knowledge graphs in large datasets. Existing ontology browsers and related knowledge-navigation platforms provide web-based interfaces for exploring structured biomedical knowledge resources[9–12]. However, most existing tools are designed for specific, well-defined concepts and relations rather than the relatively semi-structured and heterogeneous CDEs[2]. A general framework that integrates visualization with quantitative analysis of semantic structure, organizational contribution, reuse patterns, temporal growth, and redundancy examination could substantially improve CDE comprehension, development and management.

To address this gap, we developed CDEAtlas, an interactive platform that visualizes and supports exploration of CDE semantic and structural patterns. Three key requirements guided the system design: first, the need to systematically profile the CDE characteristics, second, the need for a global overview of the repository; and third, the need for multi-level exploration of the repository through semantic, organizational, and temporal perspectives[13]. CDEAtlas integrates structured statistics and semantic analysis with interactive visualization and is publicly available at https://clinicalnlp.org/cde-atlas/.

# The two authors contributed equally

* Corresponding authors: na.hong@yale.edu; huan.he@yale.edu

Methods

The design of CDEAtlas was guided by a human-centered, iterative process. Specifically, CDEs from multiple repositories are processed and transformed into semantic embeddings, followed by similarity analysis, dimensionality reduction, and clustering to generate a 2D semantic map that supports navigation, characteristic analysis, and overlap detection through the CDEAtlas interactive visualization interface. The overall workflow of the proposed method is illustrated in Figure 1, integrating these components to support both navigation of individual CDEs and broader understanding of the structure, content, and evolution of the CDE ecosystem.

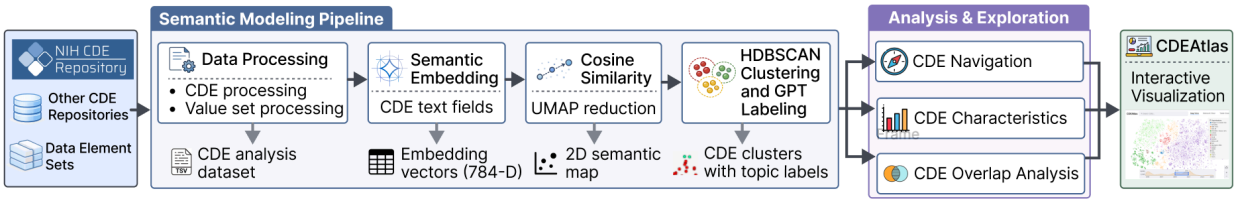

Figure 1. The overall workflow of our method

Data Source

The proposed methods are designed to be compatible with any data element resources, including those from large CDE repositories, project developed data elements across multiple studies, data dictionaries from large biomedical databases, and community generated CDEs from various research initiatives. To illustrate the core capabilities and features of CDEAtlas, we selected the NIH CDE Repository as a representative demonstration dataset, given its comprehensive coverage and the best compatibility with the NIH Data Management and Sharing Policy. The NIH CDE Repository[4] serves as a centralized catalog for CDEs submitted and endorsed by various NIH institutes and other federal agencies. Each CDE entry contains structured metadata including the question text, definition, permissible values, associated forms, and publication date. As of March 2026, the repository contains 22,743 CDEs spanning dozens of research domains, with ongoing contributions from NIH Institutes and Centers and other organizations.

Semantic Modeling Pipeline

Data processing. We retrieved 22,743 CDEs from the NIH CDE Repository via its public API and bulk JSON export. For each CDE, we extracted and structured the following fields: (1) question text and definition for semantic analysis; (2) concept annotations at the CDE level, including concept source (e.g., NCIt, LOINC, SNOMED CT) and concept identifiers; (3) value set annotations, including permissible values and their associated concept codes and sources; (4) organizational metadata, including the submitting institution and steward organization; (5) form associations, recording which clinical or research forms include each CDE; and (6) publication and last-updated timestamps for temporal analysis. Concept source availability was assessed by parsing each CDE's JSON representation, and concept frequency statistics were computed by aggregating concept identifier occurrences across the full corpus.

Semantic embedding. To characterize the semantic structure of the extracted CDE dataset, we constructed a textual representation for each CDE by concatenating the CDE name, description, and question text when available. These texts were first normalized and combined into a single string to capture the conceptual meaning of each element. The resulting texts were then encoded into dense vector representations using the google/embeddinggemma-300m[14] embedding model, producing 768-dimensional embeddings for each CDE , which represent the high-dimensional semantic similarity among the CDE dataset.

Dimensionality reduction. To enable intuitive exploration of the semantic relationships across the entire CDE dataset, the high-dimensional embeddings were further projected into a two-dimensional space using UMAP (Uniform Manifold Approximation and Projection)[15]. The UMAP were applied with n_neighbors=30, min_dist=0.5, and cosine similarity to preserve local neighborhood structures while maintaining meaningful global separation across semantic regions, allowing conceptually related CDEs to appear close to each other on the map while keeping distinct topics visually separable.

Clustering and labeling. Clustering was then performed in the resulting two-dimensional embedding space using the HDBSCAN algorithm to identify groups of varying size and density within the semantic embedding spaces of CDEs[16]. To facilitate interpretation of the discovered clusters, we adopted a cluster labeling strategy inspired by our prior work on embedding-driven clustering and LLM-based topic labeling[17]. For each cluster, we collected the textual representations of the CDEs assigned to that cluster and combined them into a single input context. This context was then submitted to the LLM (gpt-5-mini) together with a predefined prompt designed to guide the model

to summarize the dominant semantic theme of the cluster. To ensure computational efficiency and maintain representativeness, the LLM context was constructed using the full text of clusters with up to 200 members; for larger groups, a random 5% selection was implemented, with total membership restricted to a range of 200 to 800 CDEs. Following the prompt instructions, the LLM generated a text label that serves as high-level summaries that help users quickly interpret the semantic meaning of clusters in the visualization.

CDE Data Analysis & Exploration

Using the pipeline described above, we generated embeddings, clusters, and LLM-based topic labels for all CDEs. These outputs were integrated into the CDEAtlas platform to support several analyses of the semantic structure and distribution of the CDE repository.

CDE Navigation

To enable interactive exploration and navigation of the CDE semantic landscape, we developed CDEAtlas, a web-based application that integrates several visual analytics capabilities, that allows users to explore the distribution and relationships of CDEs within the semantic map derived from the embedding pipeline.

CDEAtlas incorporates a multi-modal search interface supporting keyword search across CDE metadata and property-based filtering by organization, concept source, publication date range, and form usage tier. The search module indexes CDE question text, definitions, and associated metadata to enable rapid retrieval. Upon query submission, matching CDEs are highlighted on the 2D map and listed in a sortable results panel. Each result entry provides a direct link to the corresponding CDE page on the NIH CDE Repository, facilitating seamless linkage between visual exploration and authoritative source records.

We used WebGL through the Three.js[18] JavaScript library to visualize CDEs as a point cloud on a 2D semantic map. Colors can represent attributes such as source organization, publication period, concept source, or form usage tier. The system supports real-time zoom, pan, and hover interactions. An integrated sidebar provides statistical summaries dynamically linked to the visualization, updating automatically when users select regions or apply filters.

CDE Characteristics Analyses

Beyond the semantic embedding, we further conducted five complementary analyses designed to characterize the CDE ecosystem quantitatively as follows.

(1) Concept Coverage. We quantified the usage of biomedical concept sources across the CDE repository at two levels. At the CDE level, concept sources were derived primarily from Data Element concept annotations, with Object Class annotations used as a fallback when concept annotations were absent. At the permissible value level, sources were extracted from coded permissible values containing both a concept code and a coding system. Concept sources were normalized into consistent categories (e.g., LOINC, NCI Thesaurus, and UMLS) before aggregation, and the number of CDEs annotated by the concepts from each source was counted to estimate concept coverage.

(2) Concept Frequency. We tallied the frequency of concept appearing across all CDEs (CDE-level) and across all permissible value entries (value set-level) to identify the most commonly used concepts. This analysis highlights which clinical concepts are most broadly shared across different organizations and studies.

(3) Organizational Contribution Analysis. We stratified CDEs by their steward organization to quantify each organization's contribution to the repository. We also examined the relationship between steward organizations and data sources to analyze how CDE contributions are distributed across organizations–data source combinations.

(4) CDE Usage Statistics. We analyzed the relationship between CDEs and forms by counting how many forms include each CDE. Based on form inclusion frequency, CDEs were categorized into usage tiers (≥10 forms, 5–9 forms, 2–4 forms, exactly 1 form, and not used in any form). The number and proportion of CDEs in each tier were then computed to characterize CDE use patterns across the forms.

(5) Temporal Growth Analysis. We analyzed the temporal growth of the repository by grouping CDEs according to their creation year. For each year, we computed the number of newly added CDEs and the cumulative total of CDEs in the repository. We also identified the top two steward organization(s) of CDEs contributors each year to highlight major phases of contribution. CDEs without available creation timestamps were assigned to the earliest year to maintain consistent aggregation.

Semantic Overlap Analysis for Identifying duplicate CDEs

To identify potential redundancy among CDEs, we analyzed CDE overlaps using two complementary approaches. First, exact duplicate groups were identified based on structural metadata: CDEs were grouped together if they shared the same name or an identical set of concept annotations. Concept sets were derived primarily from Data Element concept annotations, with Object Class annotations used as a fallback when concept annotations were

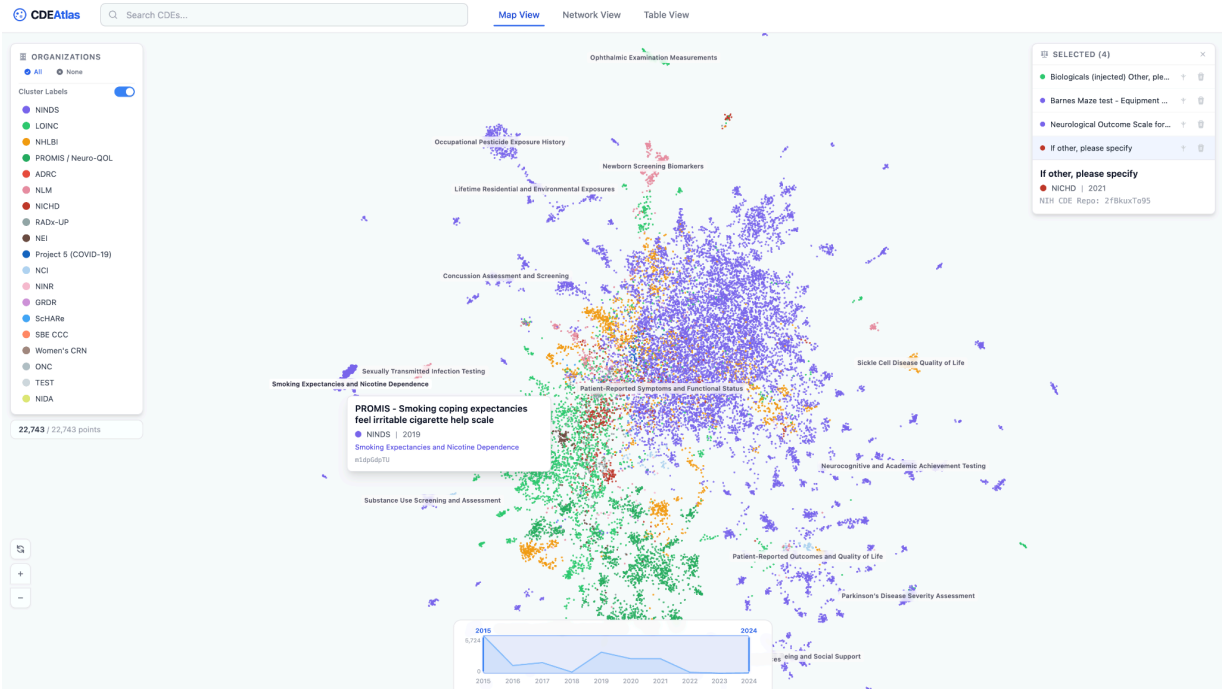

Figure 2. The landscape view visualizes all CDEs as a 2D semantic map colored by organization. Semantically related CDEs form visible clusters in the embedding space, enabling exploration of thematic concept groups and temporal trends through an integrated time-range selector and search tool.

absent. Then, highly similar CDEs were detected using semantic embeddings described above. Cosine similarity was computed between normalized embedding vectors, and CDE pairs with similarity greater than 0.95 were considered highly similar. This cutoff is consistent with common practices in semantic similarity and was chosen to maximize precision. A similarity graph was constructed and connected components were used to identify clusters of semantically related CDEs. Finally, both exact duplicate groups and highly similar clusters were summarized by group size (N = 2, N = 3, N = 4, and N ≥ 5) to characterize the scale of redundancy within the repository.

Results

CDE Navigation

Exploring the CDE landscape. All 22,743 CDEs are projected on a shared semantic space and visualized as an interactive 2D semantic landscape (Figure 2). Each element is color-coded by organization, enabling users to intuitively observe which domains were most active within the CDE repository. The integrated search bar enables efficient concept retrieval and provides direct navigation for deeper exploration within the CDE semantic space. Each dot directly links to its corresponding NIH CDE pages.

Tracking the temporal evolution of CDEs. To support longitudinal exploration, an area chart in the lower panel allows users to filter and refine the visualization by specific year ranges. This interactive feature enables the identification of temporal patterns, revealing how CDEs have evolved over time, highlighting periods of increased activity, emerging domains, or shifts in CDE development priorities across the NIH CDE Repository.

CDE Characteristics

(1) Concept Coverage

Table 1 summarizes concept coverage across the NIH CDEs at the CDE level. Concept sources were identified primarily from Data

Table 1. Concept coverage in CDEs

Coding Systems	Number of Coded CDEs
LOINC	800
NCI Thesaurus	166
NCI caDSR	125
UMLS	25
SNOMED	8
Total	1,111

Table 2. Concept coverage in value sets

Coding Systems	Number of Coded Values
LOINC	20,826
NCI Thesaurus	530
SNOMED	252
UMLS	192
ICD9CM	105
ICD10PCS	70
CDCREC	7
AdministrativeGender	3
Total	21,985

Element Concept (DEC) annotations, with Object Class used as a fallback when DEC annotation was unavailable. Counts were tabulated at the CDE level, such that each CDE was counted once according to its annotated concept source rather than by the number of distinct concept codes. Based on this procedure, only 1,111 of 22,743 CDEs (4.9%) had at least one concept annotation. LOINC was the most common source (800 CDEs), followed by NCI Thesaurus (166) and others. These findings indicate that explicit ontology mapping is available for only a small fraction of CDEs. Table 2 summarizes coding systems used for permissible value annotations. Among 72,649 permissible values, 21,985 (30.3%) were coded. LOINC accounted for most value-level annotations (20,826; 94.7%), with NCI Thesaurus, SNOMED, and UMLS appearing far less often.

(2) Most Frequently Used Concepts

As shown in Table 3, the 10 most frequently referenced concept names at the CDE level and the value level. At the CDE level, frequencies were calculated among the 1,111 CDEs with concept annotations. The most common concepts occurred only in a small proportion of annotated CDEs, with COVID-19 Infection ranked first (31, 2.79%), followed by Type (27, 2.43%) and Specify Other (23, 2.07%). Overall, the most frequent CDE-level concepts include both general semantic descriptors and domain-specific biomedical entities. At the value level, response option concepts dominate, particularly binary responses and frequency-based categories such as Yes/No and Never to Always. This pattern suggests that while CDE-level annotations are diverse and relatively sparse, value set annotations are more concentrated around standardized categorical response options.

Table 3. Top-10 most frequently occurring concepts

	Top 10 CDE-level Concepts		Top 10 Value-level Concepts	
#	Concept Name	Frequency	Concept Name	Frequency
1	**COVID-19 Infection**	31	**No**	1,079
2	**Type**	27	**Yes**	1,044
3	**Specify Other**	23	**Never**	1,014
4	**Person**	22	**Sometimes**	951
5	**Chlamydia trachomatis**	22	**Often**	738
6	**Occurrence Indicator**	20	**Always**	568
7	**SARS Coronavirus 2**	20	**Rarely**	557
8	**Neisseria gonorrhoeae**	20	**Not at all**	446
9	**Indicator**	14	**Somewhat**	385
10	**SARS-CoV-2**	13	**Quite a bit**	352

Note: CDE-level frequency counts distinct CDEs using the concept as their primary annotation. Value-level frequency counts total permissible value entries across all CDEs.

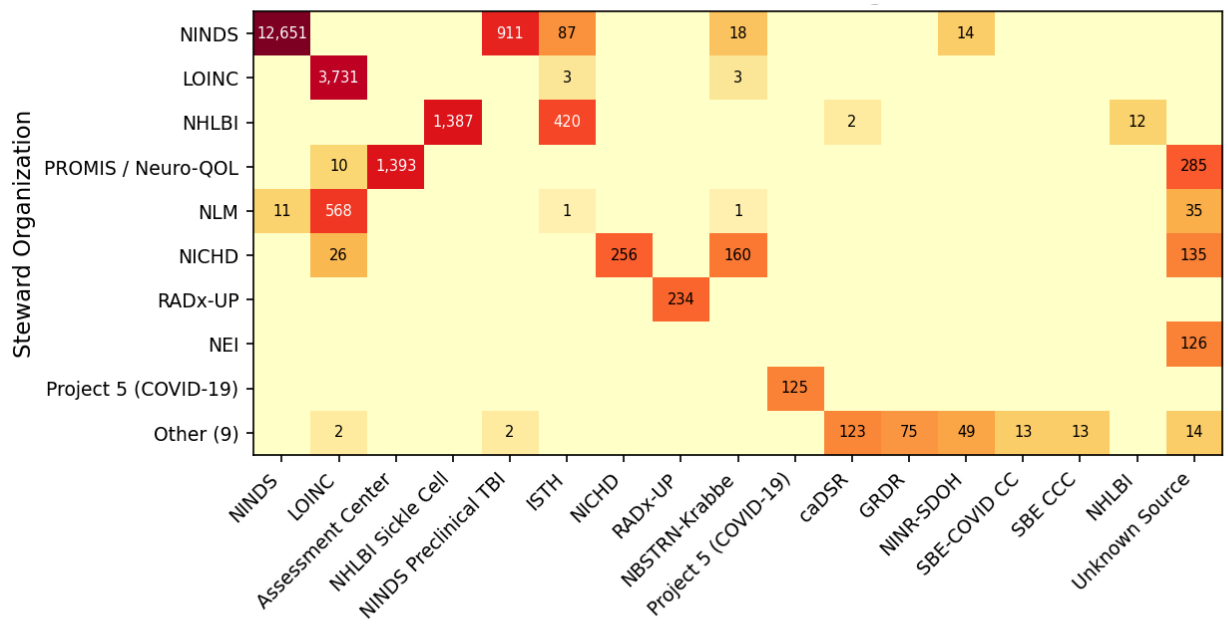

Figure 3. Heatmap showing the distribution of CDE counts across steward organizations and data sources. Rows represent steward organizations and columns represent data sources. Cell values indicate the number of CDEs contributed under each combination (sources with ≥10 CDEs shown), with color intensity reflecting log-scaled counts.

(3) Contribution by Organization

The distribution of CDEs by steward organization across the NIH CDE repository is summarized in Figure 3. NINDS is by far the dominant contributor, accounting for a total of 13,545 CDEs (59.56%) in the repository and reflecting its central role in developing standardized neurological research instruments. LOINC-linked initiatives represent the second largest contribution source, followed by NHLBI and PROMIS / Neuro-QOL, each contributing substantial numbers of CDEs across their respective research domains. The heatmap further illustrates these contribution patterns across steward organizations and data sources.

The visualization reveals that CDE contributions are highly concentrated in a small number of organization and data source combinations rather than evenly distributed across the repository. In particular, the majority of CDEs originate from a few dominant sources associated with NINDS and LOINC initiatives, while other organizations contribute smaller, more specialized sets of CDEs tied to particular projects or clinical domains.

(4) CDE Usage and Patterns

A markedly skewed reuse pattern is observed when the 22,743 CDEs are stratified by form usage tier (Figure 4). Only 233 CDEs (1.0%) are used in 10 or more forms, indicating that a very small subset of elements functions as cross-study foundational variables. In contrast, 12,283 CDEs (54.0%) appear in exactly one form, suggesting that most elements are study-specific and exhibit limited reuse. Additionally, 4,453 CDEs (19.6%) have never been incorporated into any form, representing either inactive or newly introduced elements. These patterns indicate that broadly shared CDEs tend to be reused across multiple forms but appear relatively sparsely within each form, whereas more specialized elements are concentrated within individual instruments.

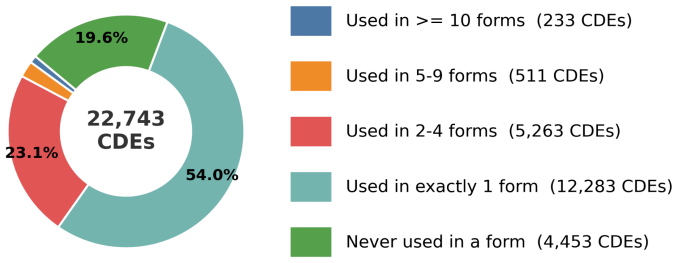

Figure 4. Distribution of CDEs by form usage frequency. Most CDEs are used in only a single form, while a smaller subset are reused across multiple forms, highlighting the uneven reuse patterns within the repository.

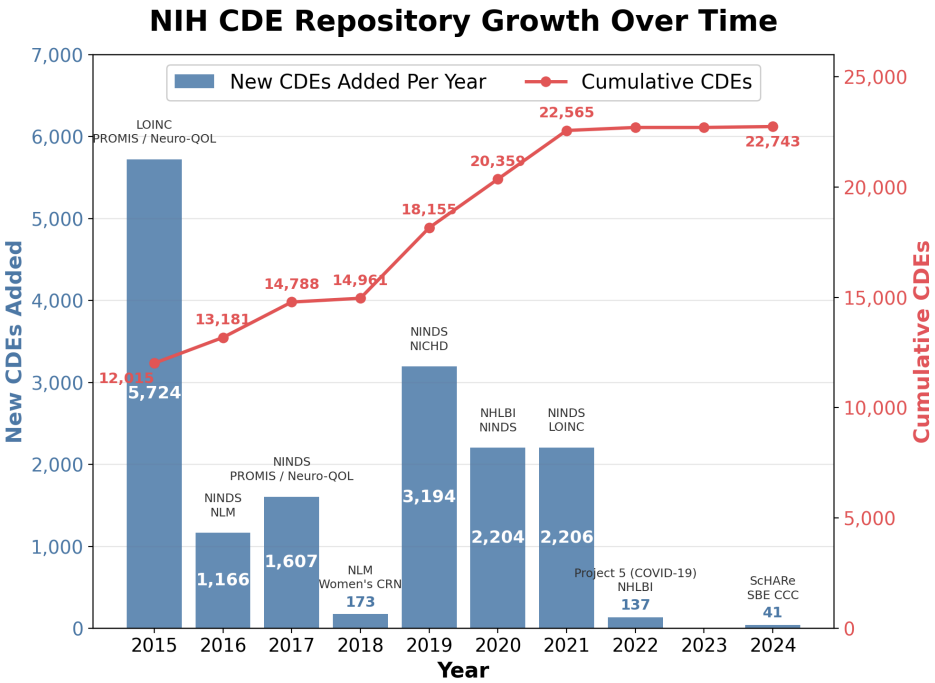

Figure 5. Temporal growth of the NIH CDE repository from 2015 to 2024. Bars show the number of new CDEs added per year, while the line indicates the cumulative total of CDEs over time.

(5) Growth Timeline of the CDE Repository

Figure 5 illustrates the temporal growth of the NIH CDE repository from 2015 to 2024. The repository expanded rapidly during the early years, with 5,724 new CDEs introduced in 2015 alone and substantial additions continuing through 2019–2021. During this period, several large-scale initiatives, particularly NINDS and LOINC-linked programs, contributed significantly to the rapid expansion of the repository.

Table 5. Distribution of exact duplicate and highly similar CDE clusters.

Overlap Tier	Exact Duplicate	Highly Similar (cosine>0.95)	Example
N>=5	26 (198)	45(400)	Which ones?; If other, please specify; Sex; Gender
N=4	13 (52)	46 (184)	Age; Grandparent birthplace; Positive affect
N=3	32 (96)	108 (324)	Employment Status; Race; City; Marital Status
N=2	408 (816)	807 (1614)	Date of Birth; Person Birth Date; Birth Weight

Note: Exact duplicates share identical names or effective concept sets. Highly similar groups are identified using cosine similarity (>0.95) over CDE embeddings, forming connected components. N denotes the size of each overlap group. For each tier, the value before parentheses indicates the number of groups, and the value in parentheses indicates the total number of CDEs.

By 2021, the cumulative number of CDEs had reached over 22,000. In 2022, we saw a cluster expansion in COVID-19 related CDEs. After 2022, the rate of new CDE additions slowed considerably, with only a small number of elements introduced in more recent years. Nevertheless, contributions continue from newer initiatives such as ScHARe and SBE CCC, suggesting ongoing but more targeted expansion of the repository.

Duplication and Semantic Overlap Among CDEs

Table 5 summarizes semantic redundancy among the 22,743 CDEs by identifying clusters of exact duplicates and highly similar elements. Exact duplicate groups are defined by identical names or identical effective concept sets, while highly similar groups are derived from cosine similarity (>0.95) over CDE embeddings. Across the repository, numerous clusters of semantically overlapping CDEs are observed, ranging from simple pairs to larger groups of five or more elements. In total, hundreds of CDEs participate in exact duplicate groups, while substantially more appear in highly similar clusters identified through embedding similarity. Larger clusters often correspond to commonly reused survey or demographic concepts such as *"If other, please specify," "Age," or "Date of Birth,"* which are repeatedly defined across different forms or steward organizations. These patterns suggest that semantically equivalent or near-equivalent CDEs are frequently created independently rather than reused. In the CDEAtlas visualization, as shown in Figure 6, these clusters appear as dense local neighborhoods in the semantic embedding space, providing a clear visual signal of redundancy and potential opportunities for harmonization across the repository.

Discussion

CDEAtlas addresses a critical gap in the current CDE management ecosystem by providing researchers, data curators, and policy makers with an intuitive and scalable platform for exploring the CDE landscape at multiple levels of granularity. By combining semantic embeddings with a statistical analysis framework, the platform transforms a previously opaque repository of 22,743 elements into a navigable and analytically interpretable knowledge space. Through coordinated visual and analytic views, users can not only locate specific elements and identify semantically related CDEs, but also examine broader structural patterns in CDE usage across organizations and over time.

Semantic Annotation Imbalance Across CDE and Value Levels

A notable finding is the substantial imbalance between concept-level and value-level semantic annotation in the NIH CDE repository. At the CDE level, only 1,111 of 22,743 elements (4.9%) contain any concept annotation, indicating that explicit ontology mapping is available for only a small fraction of CDE definitions. Stratified analysis by endorsement status and steward organization showed that concept-level annotation was highly uneven: 100% of

NIH-endorsed CDEs were annotated versus 4.1% of non-endorsed elements, and 93% of annotated CDEs came from four stewards, NLM, LOINC, NCI, and Project 5/COVID-19.

By contrast, NINDS contributed 60% of repository CDEs but no annotated CDEs, suggesting that the low overall coverage reflects steward-specific curation practices rather than a repository-wide annotation strategy. Among these annotated CDEs, LOINC is the most frequently used source, followed by NCI Thesaurus and NCI caDSR, whereas UMLS and SNOMED appear only rarely. In contrast, value-level coding is richer but remains incomplete. Among the 72,649 permissible values extracted from the repository, only 21,985 (30.3%) contain standardized concept codes, with LOINC accounting for the vast majority of coded values. This pattern suggests that semantic standardization appears to be applied more consistently to coded laboratory test measurements and results than to general concepts and response options used in survey questions. Because concept-level alignment is essential for cross-study harmonization and semantic retrieval, this gap highlights an important area for future CDE curation efforts.

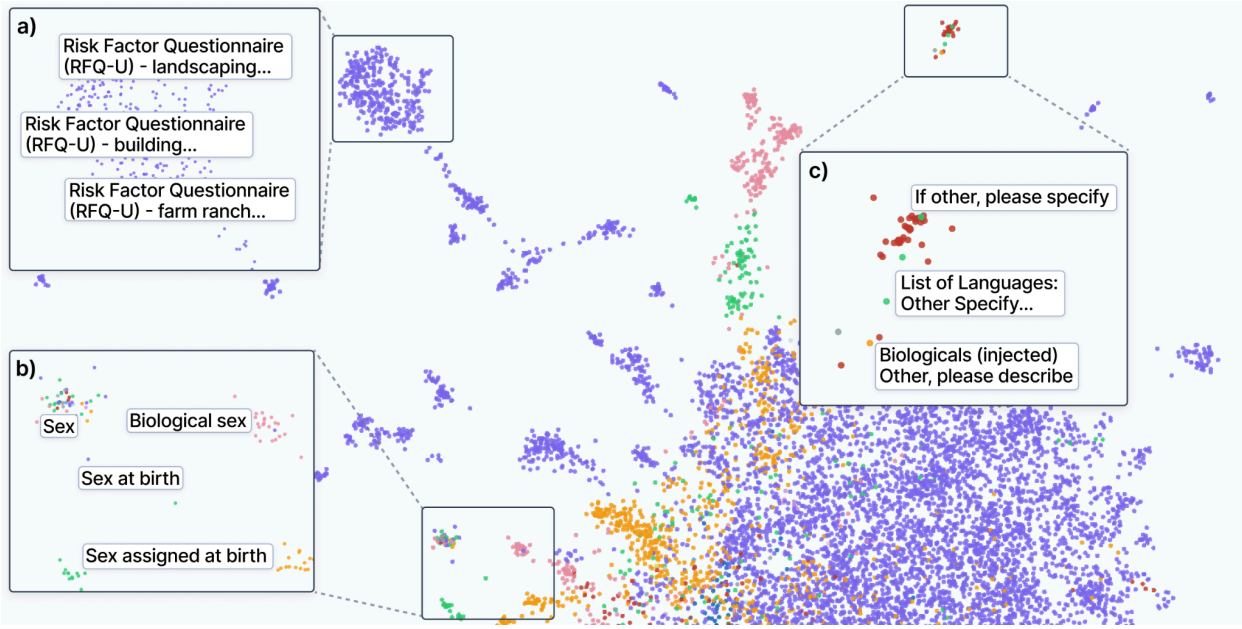

Figure 6. Representative examples of semantic overlap patterns among CDEs identified in the map, including (a) CDEs with a shared prefix but different semantic meanings, often originating from the same organization or questionnaire series; (b) CDEs with near-equivalent meanings; and (c) semantically similar CDEs contributed by different organizations.

Uneven Form Reuse Patterns in the Repository

The form usage analysis reveals a power law distribution characteristic of reuse in complex CDE ecosystems. A relatively small subset of CDEs is heavily reused across many forms, while a long tail of elements appears only once or in a limited number of forms. Approximately 20% of CDEs have no form associations in the repository, which may reflect inactive, newly introduced, or not yet reused elements. Overall, this pattern suggests that only a limited core of CDEs functions as broadly reusable building blocks across studies, whereas most elements remain tied to specific instruments or local data collection needs. Such an uneven reuse structure highlights both the presence of foundational shared variables and the continued fragmentation of CDE development and adoption across the repository.

Redundancy and Patterns of Semantic Overlap and Near-Duplicate CDEs

Redundancy across CDE definitions is also evident from both matching inspection and semantic similarity analysis. Certain foundational variables such as age, race, and sex appear in more than 7% of all CDEs, reflecting a small set of demographic elements that recur across many clinical studies. At the same time, the repository contains multiple independently defined CDEs representing nearly identical survey constructs. For example, variations of the commonly used prompt “If other, please specify” appear repeatedly across different forms and steward organizations, suggesting that functionally equivalent auxiliary variables are often recreated rather than reused from existing definitions.

This qualitative observation is further supported by the semantic overlap analysis. Approximately 8.1% of CDEs have near-duplicate counterparts with cosine similarity greater than 0.95, and more than one quarter of all elements exhibit substantial similarity to at least one other CDE. These similarity clusters often correspond to commonly reused constructs such as demographic variables or auxiliary survey prompts, highlighting opportunities for harmonization and consolidation across the repository. For example, highly similar but non-identical pairs included "*Date of Birth*" and "*Person Birth Date*," as well as questionnaire items such as "*Are you able to get up from the floor…*" versus "*Are you able to get up off the floor…*" and "*Communication device type*" versus "*Communication devices type.*"

These examples suggest that many near-duplicates arise from minor naming, wording, or grammatical variations rather than substantive semantic differences. As shown in Figure 6, semantic overlap in the repository appears in several recurring forms: clusters of CDEs that share a common prefix yet diverge semantically within the same questionnaire or organization specific series, near-equivalent CDEs that express essentially the same concept with only minor wording differences, and semantically similar CDEs contributed by different steward organizations. Together, these patterns indicate that redundancy in the repository arises not only from exact or near-exact duplication, but also from local naming conventions, repeated instrument design practices, and parallel CDE development across organizations.

Limitations and Future Directions

Several limitations should also be noted. First of all, the embedding model may not fully capture domain specific nuances, particularly for highly technical CDE definitions with limited textual context. Because our framework is fundamentally based on embedding representations, the resulting clustering and visualization are influenced by the representational capacity of the underlying embedding model. In addition, it was trained on general knowledge and may have limited exposure to emerging biomedical terminology and specialized clinical concepts. These factors may reduce its ability to distinguish subtle semantic differences among closely related CDEs.

Second, the two dimensional UMAP projection inevitably introduces distortion relative to the original high dimensional embedding space. As it is intended primarily as a visualization aid rather than a truth representation of pairwise distances in the embedding space, the UMAP preserves local neighborhood structure. Therefore, the spatial proximity in the visualization should be interpreted qualitatively rather than as an exact measure of the semantic similarity. We plan to augment the visualization by including the original high-dimensional neighbor distances between pairs to allow users to compare the projected view with the relationship in the original embedding space.

Third, the analyses and visualizations presented in this study are based on the most recent version of the available NIH CDEs at the time of analysis, reflecting the current state of NIH CDE aggregation and development. Because the NIH CDE repository is continuously updated, the findings reported here should be viewed as a temporal snapshot rather than a fixed representation of the CDE landscape. New CDE releases, revisions to existing definitions, and increased harmonization efforts across research domains may alter the semantic structure captured by the embedding model, potentially affecting downstream clustering and visualization results.

We will also enhance compatibility with mainstream data standards such as the Clinical Data Interchange Standards Consortium (CDISC)[19] and Fast Healthcare Interoperability Resources (FHIR)[20]. In addition, our platform will be dynamically updated to incorporate the latest versions of NIH CDEs, ensuring that the data remains current and representative of the evolving CDE ecosystem, and we will also incorporate open APIs to further enhance usability and adoption. Furthermore, CDEAtlas's semantic similarity framework enables it to evolve into a recommendation engine capable of suggesting existing CDEs with high similarity scores, thereby encouraging reuse and minimizing redundant element creation.

Conclusion

CDEAtlas is a novel interactive visualization and analysis platform that enables large-scale exploration of the CDEs through a visualized semantic landscape backed by a comprehensive statistical analysis framework. By characterizing concept coverage, organizational contribution, CDE use patterns, temporal growth trajectories, and semantic overlap across the entire NIH CDE Repository, CDEAtlas empowers researchers, data curators, and policy-makers to understand the CDE ecosystem at a glance, identify harmonization opportunities at different priority levels, and track the evolution of CDE development over time. As CDEs become increasingly central to biomedical research data standardization, tools like CDEAtlas will play an essential role in accelerating CDE development and management, thereby further improving the shareability and reusability of valuable research data.

References

1. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. J Biomed Inform. 2020 Jul 1;107:103421. doi:10.1016/j.jbi.2020.103421
2. Hu Z, Wang A, Duan Y, Zhou J, Hu W, Wu S. Toward Better Semantic Interoperability of Data Element Repositories in Medicine: Analysis Study. JMIR Med Inform. 2024 Sep 30;12(1):e60293. doi:10.2196/60293
3. Mayer CS, Huser V. Learning important common data elements from shared study data: The All of Us program analysis. PLOS ONE. 2023 Jul 7;18(7):e0283601. doi:10.1371/journal.pone.0283601
4. Villere S. Common Data Elements Repository. Med Ref Serv Q. 2024;43(2):182–90. doi:10.1080/02763869.2024.2323896 PubMed PMID: 38722607; PubMed Central PMCID: PMC11095837.
5. caDSR II [Internet]. [cited 2026 Mar 10]. Available from: https://cadsr.cancer.gov/onedata/Home.jsp
6. ImmPort Shared Data [Internet]. [cited 2026 Mar 10]. Available from: https://www.immport.org/shared/home
7. Wang Y, Huang J, He H, Zhang V, Zhou Y, Hao X, et al. CDEMapper: enhancing National Institutes of Health common data element use with large language models. J Am Med Inform Assoc. 2025;32(7):1130–9.
8. Tao S, Chou W, Li J, Du J, Ram P, Abeysinghe R, et al. IMI-CDE: an interactive interface for collaborative mapping of study variables to common data elements. In: 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) [Internet]. 2022 [cited 2026 Mar 10]. p. 465–8. Available from: https://ieeexplore.ieee.org/document/9874571 doi:10.1109/ICHI54592.2022.00070
9. Vendetti J, Harris NL, Dorf MV, Skrenchuk A, Caufield JH, Gonçalves RS, et al. BioPortal: an open community resource for sharing, searching, and utilizing biomedical ontologies. Nucleic Acids Res. 2025 Jul 7;53(W1):W84–94. doi:10.1093/nar/gkaf402
10. McLaughlin J, Lagrimas J, Iqbal H, Parkinson H, Harmse H. OLS4: a new Ontology Lookup Service for a growing interdisciplinary knowledge ecosystem. Bioinformatics. 2025 May 1;41(5):btaf279. doi:10.1093/bioinformatics/btaf279
11. Hong N, Wang K, Wu S, Shen F, Yao L, Jiang G. An Interactive Visualization Tool for HL7 FHIR Specification Browsing and Profiling. J Healthc Inform Res. 2019 Sep;3(3):329–44. doi:10.1007/s41666-018-0043-8 PubMed PMID: 31598581; PubMed Central PMCID: PMC6784845.
12. Cox S, Ahalt SC, Balhoff J, Bizon C, Fecho K, Kebede Y, et al. Visualization Environment for Federated Knowledge Graphs: Development of an Interactive Biomedical Query Language and Web Application Interface. JMIR Med Inform. 2020 Nov 23;8(11):e17964. doi:10.2196/17964 PubMed PMID: 33226347; PubMed Central PMCID: PMC7721550.
13. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings 1996 IEEE Symposium on Visual Languages [Internet]. 1996 [cited 2026 Mar 7]. p. 336–43. Available from: https://ieeexplore.ieee.org/document/545307 doi:10.1109/VL.1996.545307
14. Vera HS, Dua S, Zhang B, Salz D, Mullins R, Panyam SR, et al. EmbeddingGemma: Powerful and Lightweight Text Representations [Internet]. arXiv; 2025 [cited 2026 Mar 10]. Available from: http://arxiv.org/abs/2509.20354 doi:10.48550/arXiv.2509.20354
15. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv; 2020 [cited 2026 Mar 10]. Available from: http://arxiv.org/abs/1802.03426 doi:10.48550/arXiv.1802.03426
16. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. J Open Source Softw. 2017 Mar 21;2:205. doi:10.21105/joss.00205
17. Chang CH, Ondov B, Choi B, Peng X, He H, Xu H. TopicForest: embedding-driven hierarchical clustering and labeling for biomedical literature. J Biomed Inform. 2025 Dec 1;172:104958. doi:10.1016/j.jbi.2025.104958
18. Angel E, Haines E. An interactive introduction to WEBGL and three.JS. In: ACM SIGGRAPH 2017 Courses [Internet]. New York, NY, USA: Association for Computing Machinery; 2017 [cited 2026 Mar 10]. p. 1–95. (SIGGRAPH '17). Available from: https://dl.acm.org/doi/10.1145/3084873.3084875 doi:10.1145/3084873.3084875
19. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. J Biomed Inform. 2016 Apr;60:352–62. doi:10.1016/j.jbi.2016.02.016 PubMed PMID: 26944737; PubMed Central PMCID: PMC4837012.
20. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform. 2022 Jul 19;10(7):e35724. doi:10.2196/35724 PubMed PMID: 35852842; PubMed Central PMCID: PMC9346559.