Centre for Social Informatics Working Paper Series

Semantic Search and Question Similarity in Social Science Data Archives: An Overview of Institutional Implementations, Research Tools, and Technical Approaches

Vasja Vehovar and Gregor Čehovin

Centre for Social Informatics
Faculty of Social Sciences, University of Ljubljana
www.cdi.si

March 2026

Introduction

This overview examines semantic search and question similarity methods as applied in social science data archives and survey research infrastructure. Three complementary perspectives are presented. First, institutional implementations at major data archives and question banks (including CESSDA EQB, GESIS, ICPSR, ESS, and CDC Q-Bank) are reviewed. Most rely on keyword and controlled vocabulary indexing, with only limited adoption of semantic search capabilities. Second, emerging research tools (Harmony, SBERT-LaBSE, and Semantic Search Helper) are catalogued. These employ transformer-based sentence embeddings (SBERT, LaBSE, OpenAI ADA) to enable similar-wording search across survey instruments, including in cross-lingual settings. Third, the underlying technical approaches are organized into four categories: vector embeddings for semantic similarity, question similarity metrics (e.g., BERTScore, SimCSE), domain-specific efforts at survey organizations, and hybrid methods combining keyword matching with embedding-based retrieval.

The overview indicates that the technical foundations for semantic question search are well established; however, their integration into operational data archive infrastructure remains at an early stage. Closer integration of semantic search capabilities into existing data archive infrastructure remains a key area for further development, as current implementations have yet to fully exploit the available methodological foundations.

Table 1: Institutional pilots in question similarity

Organization	Objectives	Standards	Status	Similar-wording search?	Model(s) used
CESSDA — European Question Bank (EQB)	Cross-national question bank to search/browse/compare questions across European archives; multilingual support and filters.	DDI-Lifecycle; metadata schema documented by GESIS EQB Metadata Schema	Public beta (~3.5k questions); EQB Tool Page	No keyword/metadata search only	Not stated; DDI-Lifecycle metadata
GESIS — Semantic Question Annotation & Knowledge Graph	ML-based annotation of survey questions into features; exposed as a knowledge graph and integrated into semantic search.	SEMANTiCS 2020 Paper; GESIS KG Documentation	Prototype (~4k questions); SPARQL demo available	Partial semantic »facets« (not full similar-wording)	Neural classifiers for question features; Knowledge Graph
ICPSR — Social Science Variables Database (SSVD)	Variable/question-level discovery across studies; search by question text, value labels, and variable labels.	Structured metadata indexing; integrated into ICPSR discovery stack	Operational and expanding; Find Data Portal	No keyword/metadata search only	Not stated; standard text indexing
ESS — Data & Metadata Portal	Cloud-hosted data/metadata service with FAIR-compliant metadata, APIs, and powerful search for questions/variables.	FAIR principles; DDI metadata; SSHOC/ESS Service	Work in progress; ESS Portal	No keyword/metadata search only	Not stated
CDC / NCHS — Q-Bank	Repository for question evaluation (cognitive testing, methods reports); links questions to NHIS survey data.	Evaluation metadata; CCQDER Program	Operational and updated regularly	No keyword/metadata search only	Not stated
CESSDA — Data Catalogue	Multilingual keyword filtering and controlled vocabularies for harmonized question discovery across archives.	Controlled vocabularies; multilingual indexing; FSD News	Live service; Tool Page	No keyword/metadata search only	Controlled vocabulary & metadata

Table 2: Research tools for question similarity

Tool	Function	Methods	Maturity	Status	Similar-wording search?	Model(s) used
Harmony	Helps harmonize questionnaire items by finding semantically similar questions across instruments.	Transformer embeddings; cosine similarity; packaged workflows	Live app + code; validated in BMC Psychiatry 2024	Ready-to-use for deduplication and harmonization; complements institutional search tools	Yes embedding-based similar-wording search	Transformer-based (SBERT)
SBERT-LaBSE (Kang et al.)	Cross-lingual detection of redundant health survey questions (English–Korean); builds STS dataset and evaluates models.	SBERT + LaBSE embeddings; cosine similarity.	Published in JMIR Medical Informatics 2025; DOI: 10.2196/71687	SBERT-LaBSE (Kang et al.)	Yes embedding-based similar-wording search	Transformer-based (SBERT + LaBSE)
Semantic Search Helper	Prototype to harmonize cohort items at scale; surfaces candidate pairs of semantically similar questions.	SBERT or OpenAI ADA embeddings; interactive filtering and visualization.	Feasibility study in European Psychiatry 2025; DOI: 10.1192/j.eurpsy.2024.1808; GitHub Prototype	Semantic Search Helper	Yes embedding-based similar-wording search	Transformer-based (SBERT + OpenAI ADA)

Table 3: Approaches to semantic similarity

Maturity	Organizations	Papers (links)	Contacts	Similar-wording search?	Model(s) used
Approach 1: Vector Embeddings for Semantic Similarity					
Fully mature; widely used in production for semantic search and question retrieval	• OpenAI (San Francisco) • Hugging Face (New York) • Microsoft (Redmond)	• VectorSearch: Enhancing Document Retrieval with Semantic Embeddings — Microsoft Research • A Comparative Study of Sentence Embedding Models for Assessing Semantic Variation — Hugging Face	OpenAI: support@openai.com Hugging Face: contact@huggingface.co	Yes embedding-based similar-wording search	Transformer-based (e.g., SBERT)
Approach 2: Question Similarity Metrics					
Mostly research stage; used in evaluation and some QA systems	• U.S. Census Bureau (U.S. Department of Commerce) • NCES = National Center for Education Statistics (U.S. Department of Education)	• BERTScore: Evaluating Text Generation with BERT — Columbia University • Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning — TU Darmstadt	Federal Committee on Statistical Methodology	Partial Primarily for evaluation; not typically exposed as end-user similar-wording search	BERTScore / cross-encoder similarity / unsupervised sentence embeddings (e.g., SimCSE)
Approach 3: Domain-Specific Efforts					
Active pilots in survey organizations; semantic	• CESSDA (Bergen, Norway)	• Semantic Annotation, Representation and Linking of Survey Data — GESIS	CESSDA: cessda.net/contact ICPSR: icpsr-help@umich.edu	Partial emerging feature-based semantic facets; most	Domain ML (e.g., LSTM classifiers for features at

search in question banks is emerging	• ICPSR (University of Michigan) • GESIS (Germany)	• A Systematic and Comparative Analysis of Semantic Search Algorithms — VishwakarmaInstitute of Information Technology		archival portals lack similar-wording UI	GESIS); many archives rely on keyword / ELSST thesaurus
Approach 4: Hybrid Approaches (Keyword + Semantic)					
Becoming best practice in enterprise search and Retrieval-Augmented Generation (RAG)	• Elastic, Elasticsearch (Mountain View) • Adobe Research (San Jose)	• Hybrid Semantic Search: Unveiling User Intent Beyond Keywords — Innoplexus Consulting Services • Hybrid Retrieval for Hallucination Mitigation in LLMs — Pisa & SNS	Elastic: info@elastic.co Adobe: research@adobe.com	Yes hybrid stacks combine keyword with embedding similarity	Transformer-based (e.g., SBERT)