

Article

A Conceptual Framework for Semantic Indexing of Data Sources Based on Structured Peer-to-Peer Model, Hilbert Curve, Hypercube and Data Analysis

Mohammed Ammari ^{1,*} , Fadwa Ammari ²  and Abdelaziz Boumahdi ² 

¹ Faculty of Sciences, Mohammed V University, Rabat 10080, Morocco

² Faculty of Educational Sciences, Mohammed V University, Rabat 10080, Morocco; fadwa.ammari@um5r.ac.ma (F.A.); a.boumahdi@um5r.ac.ma (A.B.)

* Correspondence: m.ammari@um5r.ac.ma

Abstract

Semantic indexing ensures better organization and optimized searching of heterogeneous, autonomous, and distributed data sources. This approach leverages meaning and context rather than just keywords to better manage the increasing volume, complexity, and heterogeneity of modern data, enabling precise searching, optimized integration, and improved interoperability between domains. Several approaches to semantic indexing are available: ontology-based indexing, machine learning and automated semantic annotation of data sources. However, the main challenge remains scaling up. This article focuses on a conceptual framework designed for scalable semantic indexing of data sources based on a structured peer-to-peer architecture adapted for managing a very large number of nodes, Hilbert curve renowned for its preservation of semantic affinity while scaling, hypercube structure with its efficient diffusion algorithm, semantic annotation of data sources based on keywords, as well as machine learning techniques, in particular, multidimensional data analysis. An illustrative exploratory example of the *Meta Skills* semantic class is presented to outline the proposed architecture. This study proposes a conceptual and exploratory framework for large-scale semantic indexing of data sources. The proposed approach has not yet been implemented or validated on a large scale; its objective is to provide an initial structured model to serve as a basis for future empirical research.

Keywords: semantic indexing; data integration; peer-to-peer system; Hilbert curve; hypercube; data analysis; Meta Skills



Academic Editor: Kesheng (John) Wu

Received: 4 March 2026

Revised: 26 April 2026

Accepted: 29 April 2026

Published: 5 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Semantic indexing of data sources represents a crucial advance in information management and exploitation, transcending the limitations of keyword-based approaches. Assigning contextual meaning to data contributes to a deeper understanding and relevant retrieval of information. Indeed, semantic indexing is essential for modern information systems, offering increased capacity to respond to complex queries, automate data analysis and support informed decision-making in increasingly vast and diverse data environments. Several approaches to semantic indexing are available:

- **Ontology-based indexing:** Many methods rely on the use of ontologies to ensure reconciliation, integration and mapping of distributed, autonomous and heterogeneous data sources. Indeed, ontology, which takes the form of a conceptual schema

containing a structured set of concepts, provides an effective mechanism to guarantee unified access, optimized query reformulation, and better semantic reconciliation of disparate data sources (databases, websites, sensor data, etc.) [1,2].

- Machine learning and lexical embeddings: Several recent methods rely on machine learning, knowledge graph lexical embeddings, and neural lexical embeddings to ensure integrated semantic indexing, as well as improved efficiency and precise searching based on the detection of latent relationships not directly uncovered [3].
- Automated semantic annotation of data sources: This is an approach that links the inherent attributes of data sources with concepts extracted from a semantic domain. This results in a marked improvement in the quality and efficiency of research, as well as increased interoperability between domains [4,5].

Certainly, the semantic indexing approaches presented earlier make it possible to improve the efficiency of searching for relevant resources by optimizing network traffic performance. However, the main obstacle remains the scaling-up factor. To overcome this obstacle, one possible solution would be to opt for an architecture based on the peer-to-peer paradigm, in this case the structured peer-to-peer mode based on a distributed hash table (DHT) (Chord [6], CAN, Tapestry, Pastry, etc.). Indeed, the structured P2P DHT mode offers the possibility of a simple, efficient and reliable resource search (example: any data can be accessed in $O(\log(n))$ hops with an underlying Chord network with n participants), and displays good performance in terms of scalability and fault tolerance. On the other hand, this mode is rather suited for unidimensional search (with one search attribute) and exact match search, also suffers from an intrinsic semantic problem.

This study proposes a conceptual and exploratory framework combining the three approaches mentioned above, as well as the structured P2P model. The proposed structure relies on the use of ontologies, machine learning and semantic annotation of data sources. The proposed framework is built on a structured peer-to-peer system suitable for scaling, Hilbert curve renowned for its preservation of semantic affinity, hypercube structure with its efficient diffusion algorithm, semantic annotation of data sources based on keywords, as well as machine learning techniques, in particular, multidimensional data analysis. This work is positioned as a conceptual and exploratory contribution. It does not aim to provide a detailed validation of the proposed architecture, but rather to propose a structured analytical framework that can guide future empirical and computational implementations.

2. Related Work

Ref. [7] propose WISE (Web-Intelligent Semantic Extractor), an intelligent framework based on deep learning, designed to extract contextually relevant information from dynamic newspaper databases by integrating natural language processing (NLP) and neural networks in order to improve both accuracy and efficiency in data retrieval. The system dynamically adjusts its crawling strategies based on content semantics, learning from patterns across diverse data sources to increase relevance and reduce noise. WISE would benefit from exploring real-world industrial applications involving large datasets in order to refine its scalability, performance and adaptability to various web ecosystems.

Ref. [8] propose a new real-time semantic indexing (RTSI) model that aims to improve information retrieval and analysis capabilities by integrating semantics into the indexing process during data ingestion. Contextual meaning is assigned to data items in real time using lightweight natural language processing (NLP), entity recognition, topic modeling, and knowledge integration. The work would benefit from being expanded by enhancing RTSI's functional capabilities to cover a wider range of events, redesigning it to handle massive volumes of data, and optimising its structure.

Ref. [9] examine how LLMs transform the integration of heterogeneous data. It explores data textualization techniques, bi/cross-encoder architectures, and the use of RAG (Retrieval-Augmented Generation) to automate semantic joins and entity resolution. However, in order to fully realise the potential of this technology, substantial challenges relating to the underlying ethical considerations must be overcome.

Ref. [10] propose Taxonomy-guided Semantic Indexing (TaxoIndex) framework to enable effective academic concept matching for paper search. TaxoIndex extracts key concepts from papers and organizes them as a semantic index guided by an academic taxonomy, and then leverages this index as foundational knowledge to identify academic concepts and link queries and documents. The proposed system could improve in terms of accuracy, scalability and long-term relevance if it supports the refinement and updating of the semantic taxonomy used.

To better position the proposed framework within the existing literature, a structured comparison of related approaches is provided in Table 1.

Table 1. Comparison between different approaches to semantic indexing.

	WISE	RTSI	TaxoIndex
Architectural adaptability	+	++	++
Evaluation method	++	+	++
Scalability mechanism	—	—	—
Semantic structuring	+	+	+ / —

“++” indicates strong support or full implementation; “+” indicates partial support; “—” indicates absence; “+ / —” indicates limited or conditional support based on specific assumptions.

This comparison highlights the absence of integrated conceptual frameworks combining semantic structuring and exploratory analytical modeling suitable for extended scaling, a gap that the present study aims to fill conceptually.

3. Methodological Innovation and Integration Logic

3.1. Research Design

This study adopts a qualitative–analytical research design grounded in data analysis and semantic indexing of data sources. The research is structured to ensure transparency, reproducibility, and conceptual rigor. The approach is exploratory and explanatory; it aims to analyze how semantic indexing mechanisms contribute to better information dissemination in distributed environments. To ensure the design of our system, we adopted the following research methodology (see Figure 1):

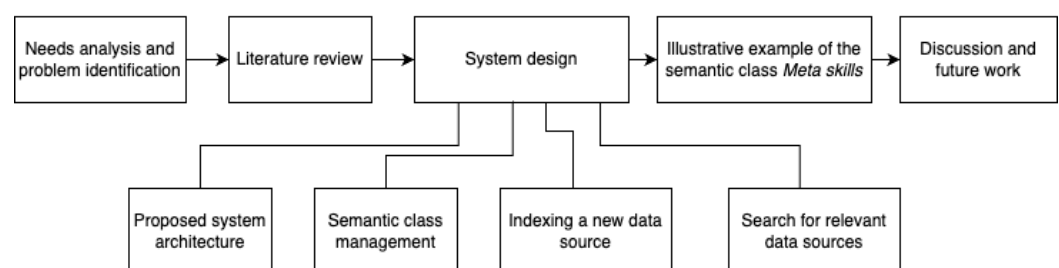


Figure 1. Research methodology.

3.2. Proposed System Architecture

The system is subdivided into semantic classes. A bijection is established between subdivisions resulting from Hilbert curve construction and semantic classes extracted from a global ontology (see Figure 2). The ontology works in a tree structure, dividing the field of knowledge into classes (Class 1: Mathematical Sciences, class 2: Life sciences,

class 3: Earth sciences, class 4: Physics, class 5: Chemistry, class 6: Legal sciences, class 7: Economics, class 8: Social sciences, class 9: Arts, class 10: Culture, class 11: Education, class 12: Information Technology, class 13: Artificial Intelligence, class 14: Cybersecurity, class 15: Data science, class 16: Engineering sciences), which are themselves divided into nested subclasses. It allows us to define, for each data source, according to a certain degree of precision, a series of numbers called an index that corresponds to its subject. The semantic classes are extracted from the *Universal Decimal Classification UDC* [11] mainly used in university and technical libraries. Several semantic classes have been added to support the rapid development of new information and communication technologies such as: Information Technology, Artificial Intelligence, Cybersecurity and Data Science

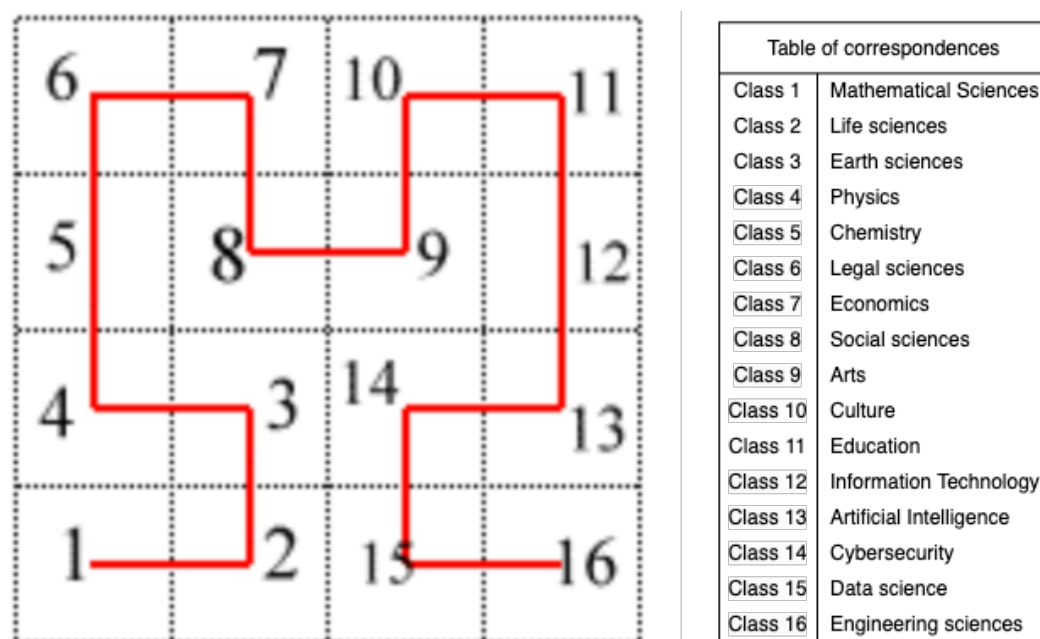


Figure 2. Correspondence between Hilbert zones and semantic classes.

Each semantic class is responsible for storing, managing and indexing metadata related of data sources based on its semantics. A semantic class assembles records that exhibit a strong semantic affinity and a weak semantic coupling with other semantic classes. A semantic class can contain several peers in order to ensure information backup and to parallelize the processing of analyses performed. In addition, nodes belonging to a semantic class store addresses of nodes, scattered throughout the network, that possess a resource expressible by concepts inherent to that semantic class. Peers belonging to a semantic class are organized in the form of hypercube architecture (see Figure 3).

3.3. Workflow Within the Semantic Class

The proposed workflow proceeds as follows (see Figure 4):

This pipeline ensures that semantic indexing decisions are directly grounded in statistically derived semantic structures rather than heuristic or purely lexical similarities.

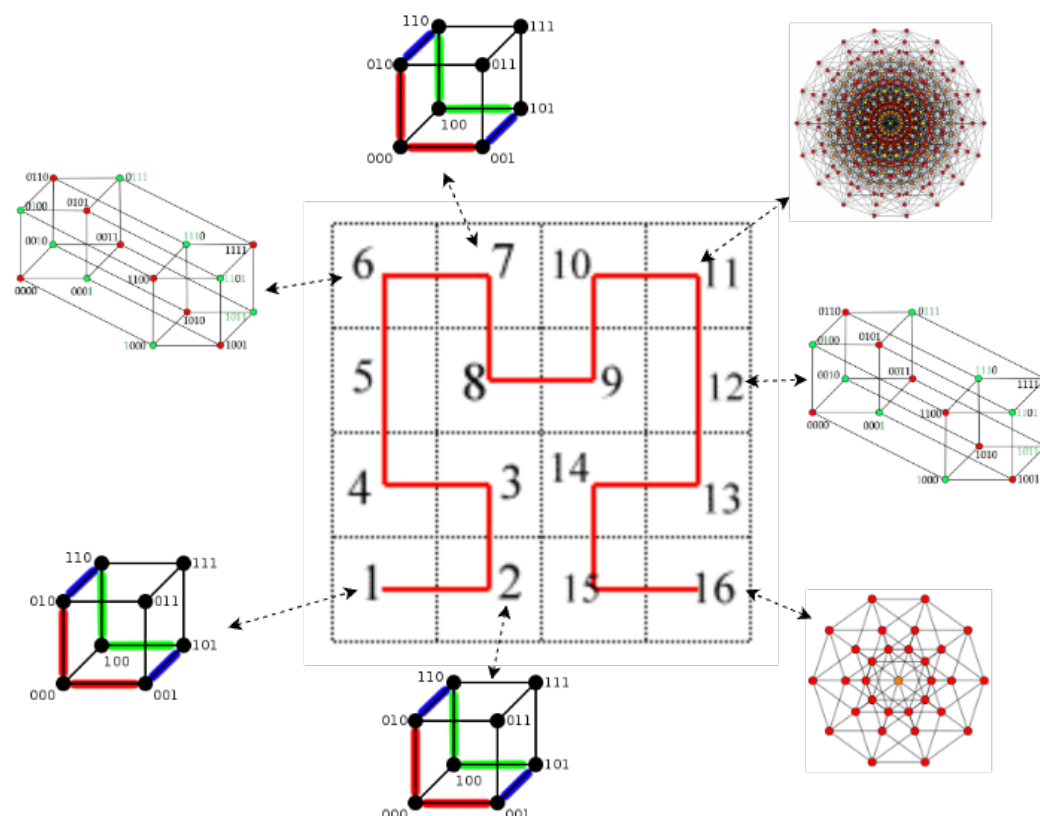


Figure 3. Overview of the proposed system.

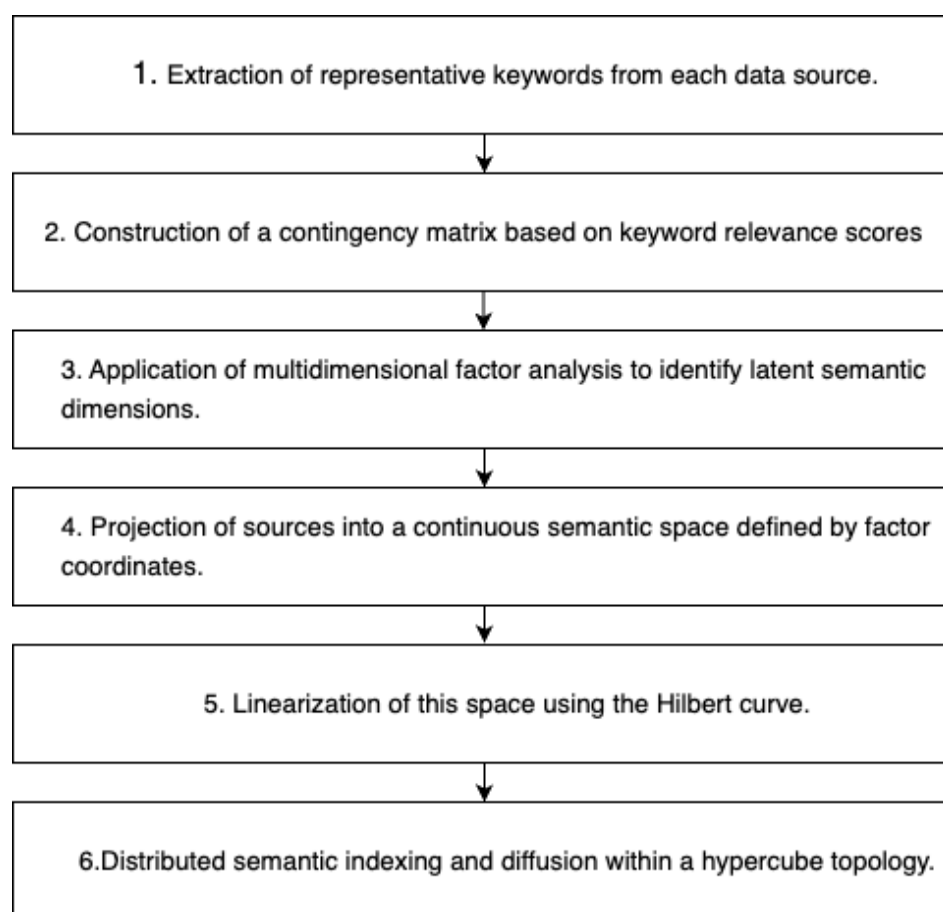


Figure 4. Workflow within the semantic class.

4. Justification of Structured and Topological Choices

4.1. Choice of Hilbert Curve

Space-filling curves are curves defined recursively in multidimensional space and whose plotting tends towards a complete covering of space as the order of the curve tends towards infinity. They allow for a bijective transformation between a D-dimensional space and a one-dimensional space (curvilinear coordinate of a point). Two neighboring points on the curve are close in space. Conversely, not all points that are close in space are systematically close on the curve. Space-filling curves are applied in various fields to preserve locality when linearizing multidimensional data (databases, signal and image processing, scientific visualization, scientific computing, etc.) [12].

The Hilbert curve is a fractal curve based on a simple pattern that takes the form of a U which can undergo rotation and reflection depending on the quadrant in which it is refined (see Figure 5). These transformations make it possible to reduce the jump of the curve from one quadrant to another. The recursive construction, combined with various rotations and symmetries, gives the Hilbert curve its properties of preserving locality, even when the number of dimensions is very high.

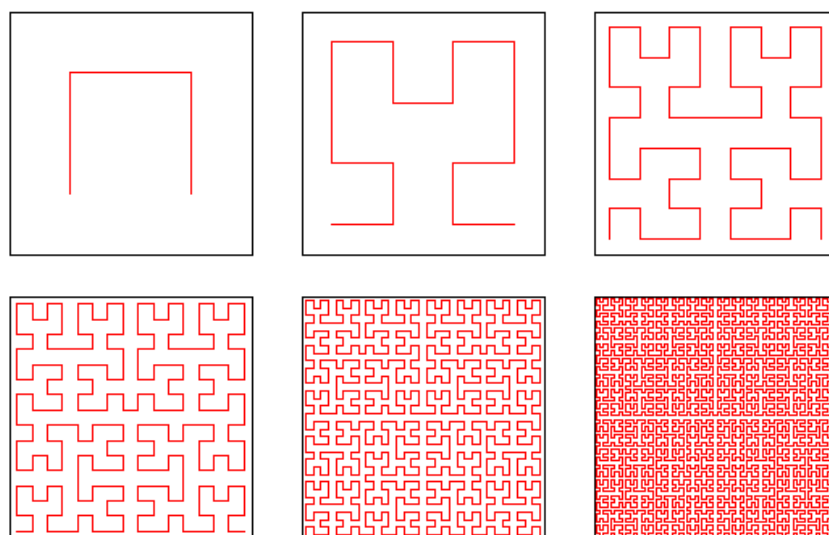


Figure 5. Recursive construction of the Hilbert curve.

The Hilbert curve can be efficiently expressed using a Lindenmayer system or a rewrite system (L-system) [13].

Alphabet: A, B

Constants: $F, +, -$

Axiom: A

Production rules:

$$A \rightarrow -BF + AFA + FB- \quad (1)$$

$$B \rightarrow +AF - BFB - FA+ \quad (2)$$

Here, F means *draw forward*, $-$ means *turn left* 90° , $+$ means *turn right* 90° , and A and B are ignored during drawing.

The choice of Hilbert curve is justified by its superior linearisation of the ontology's tree structure while preserving locality and semantic affinity. This is confirmed by comparing the Hilbert curve with other fractal curves, for example, the Lebesgue curve or z-curve, which initially uses a basic motif in the form of the letter Z repeated recursively without transformation and while maintaining the same orientation, this forces the z-curve to make sharp jumps that undeniably penalize the locality of the elements.

4.2. Choice of Peer-to-Peer System

Peer-to-peer system is a potential solution to overcome the inherent limitations of client/server model. In contrast to the latter, which uses a centralized structure based on a clearly identified server, where clients access data by establishing a connection to that server, a peer network lacks centralized coordination. Each peer acts as both client and server, offering and consuming “services” (see Figure 6). Scalability is a key factor in P2P mode. This helps to overcome the saturation phenomenon observed in C/S mode when the number of clients connecting to a server to access data becomes significant. The concept of autonomy is extremely important in P2P systems. Indeed, the autonomy given to peers allows them to control the information they possess and allows them to freely join or leave the system.

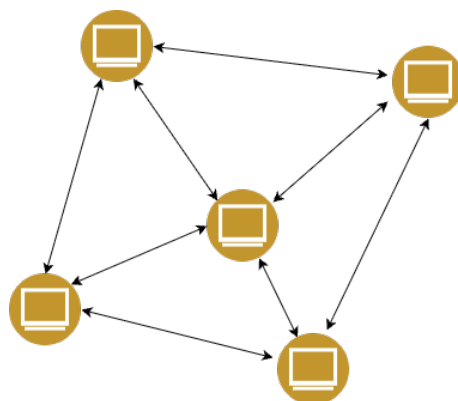


Figure 6. Peer-to-peer system.

4.3. Choice of Hypercube Structure

The hypercube is of cardinal interest in both discrete mathematics and computer science. This interest is justified by the use of its structure in many application areas (parallel architectures, information transfer, multi-criteria decision-making, coding theory, etc.). Hypercube is an n -dimensional graph generalizing notion of cube. A hypercube of dimension n has as its set of points (vertices) the binary words of length n and there exists a relation (edge) between two points if and only if the corresponding words differ in only one component. Each vertex is connected to exactly n neighbors. This means that every vertex has n degrees; in other words, the graph is n -regular [14] (see Figure 7).

The choice of the hypercube architecture is justified by the efficiency of its diffusion algorithm. An asymptotic complexity study demonstrates that the hypercube offers the best possible complexity, which is the logarithmic complexity $O(\log(n))$. It is important to note that the diffusion algorithm proposed by [15] always assumes that the hypercube is complete. In the opposite case of incompleteness (for example, if it only has seven nodes), the algorithm adds enough dummy nodes to make it complete.

Ref. [15] proposes an efficient broadcast algorithm based on the architecture of an N -dimensional hypercube. This algorithm, with logarithmic complexity, guarantees that each node receives the broadcast message only once, without duplication. Indeed, $\log_2(N)$ transmission steps are required to route the broadcast message to all nodes. We present below an illustrative example of the proposed algorithm (see Figure 8). First, node A broadcasts a message to its three neighbors: nodes E, B, and C, marking the edges on which the message was sent. The nodes receiving the message limit its transmission to edges labeled with higher numbers. Thus, node E, which received the message on a level 0 edge, transmits it only to its level 1 neighbor, node G, and its level 2 neighbor, node F. Simultaneously, node C, which received the broadcast message across a level 1 edge, transmits it to its level 2 neighbor, node D. During the third and final transmission, node G relays the message to its level 2 neighbor, node H.

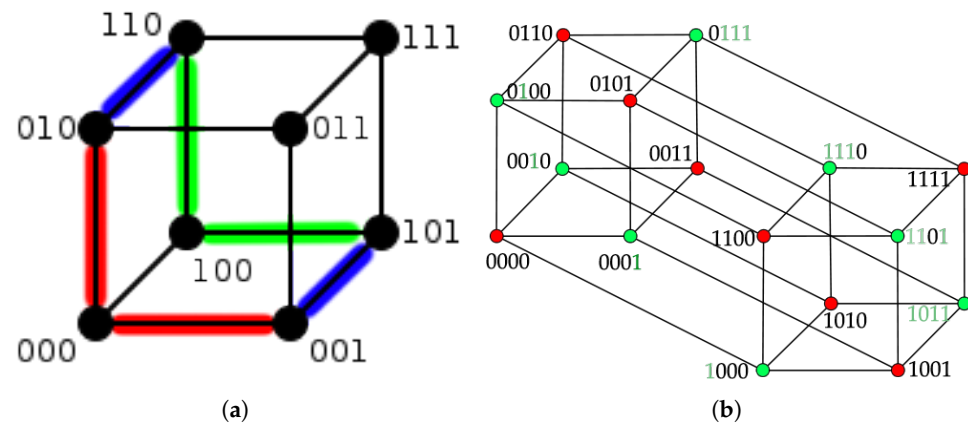


Figure 7. (a) Hypercube of dimension $n = 3$. (b) Hypercube of dimension $n = 4$.

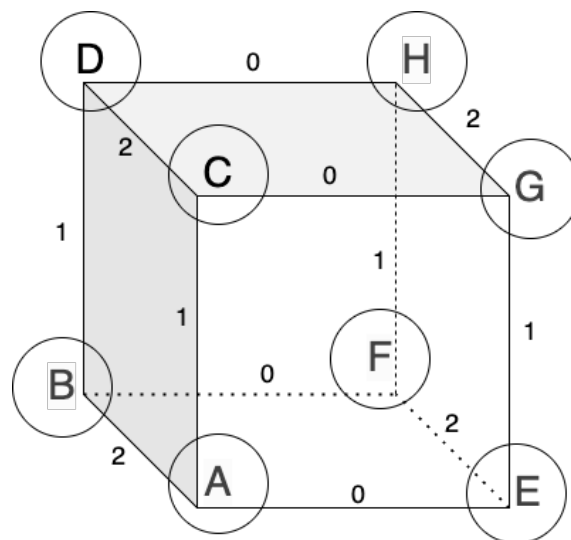


Figure 8. Three-dimensional hypercube.

The following peer management mechanisms are presented as a conceptual design intended to illustrate the feasibility of the proposed architecture, rather than as a fully implemented system.

Ephemeral peer management plays a central role in the dynamic management of the hypercube, where frequent departures (high churn) can affect the stability of the underlying routing structures as well as resource availability. One initial approach is to explicitly model churn by distinguishing short-lived peers with limited performance from more reliable and stable peers, and then adapting their presence mechanisms (for example via dynamically adjusted TTL (Time To Live) values and prioritization policies granted to long-lived peers) in order to limit the impact of unpredictable departures. A second approach relies on the concept of elderly or stable peers, which act as stabilizers of the hypercube structure by ensuring continuity of routing and storage responsibilities, thanks to their high reputation or seniority. These peers have priority for building the hypercube structure and thus play a reliable base role in redistributing the obligations associated with departing peers, limiting the impact of sporadic departures on connectivity and overall system performance and ensuring the updating of the hypercube structure. To this end, monitoring the presence of peer members within the hypercube structure using liveness mechanisms (periodic presence messages or heartbeats broadcast within the hypercube) enables the detection of silent disconnections: the absence of a response beyond a certain threshold causes neighbors to mark the peer member as inactive, remove it from the routing table and automatically redistribute its associated

responsibilities, which helps to maintain the consistency, availability and resilience of the hypercube structure in the face of high churn.

The mechanisms described in this subsection (e.g., TTL-based policies, stable peer prioritization, and heartbeat mechanisms) are presented at a conceptual level and are not implemented within the current framework. These elements should be considered as design perspectives for future implementation, rather than as operational components of the present study.

5. Semantic Class Management

5.1. Construction of the Contingency Matrix

Each semantic class acts as a container for metadata relating to data sources that have semantic affinities with that class. In practical terms, each semantic class manages a contingency matrix that ensures correspondence between I , the set of data sources, and J , the set of relevant terms related to the semantic domain. The contingency matrix $(f_{ij})_{(i,j) \in [1,n] \times [1,p]}$ represents occurrence frequency of extracted term T_j in data source S_i with $n = \text{Cardinal}(I)$ and $p = \text{Cardinal}(J)$.

Managing the contingency matrix presents enormous challenges given its size and the dispersion of its data. To overcome this obstacle, we propose using multidimensional data analysis to subdivide each semantic class into smaller, specialized subclasses. This results in increased scalability and a significant reduction in complexity. In this article, we opted for the use of factor analysis. This is a statistical method used to reduce the dimensionality of data and identify underlying factors that explain the correlations between variables. This method is applied in several fields, including marketing, psychology, social sciences, scientific research, and complex data analysis, etc. We propose the following inputs for the factor analysis method:

- A scatter plot $N(I) = \{X_i \in R^p, i \in I\}$. Each data source S_i is associated with the vector $X_i = (x_{ij})_{j \in [1,p]}$ defined by $x_{ij} = \frac{f_{ij}}{f_{i.}}$ with f_{ij} represents the frequency of the term T_j in the data source S_i and $f_{i.}$ represents the frequency of S_i in the entire population.
- A weight assigned to point X_i . We associate the weight $m_i = f_{i.} = \sum_{j=1}^p f_{ij}$ with the point X_i
- A metric that calculates the similarity between points $X_{i \in [1,n]}$. We use the distance χ^2 defined by:

$$\chi^2(X_i, X_{i'}) = \sum_{j \in J} \left(\frac{1}{f_{j.}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \right) \quad (3)$$

For Output:

- The axis of maximum elongation of the scatter plot $N(I)$ corresponds to the first eigenvector corresponding to the largest eigenvalue of the matrix of inertia $V = X^t M X$ where X is the matrix of initial data centered by the centroid G , and M is the diagonal matrix of weight.
- Factorial analysis consists of finding axes of inertia and projecting points X_i onto the first factorial plane formed by the first two eigenvectors.
- A set of decision-making indicators are being considered in order to assess the quality of the analysis performed.

5.2. Illustrative Example of the Meta Skills Semantic Class

This study adopts a qualitative-analytical research design grounded in documentary analysis and thematic synthesis. To illustrate and explain our proposals, we chose *Meta Skills* as a semantic class. We selected 18 relevant, representative, reliable, accessible

and heterogeneous data sources (see Table A1) covering the chosen semantic class and considered the following relevant candidate keywords: Knowledge K, Memory M, Problem-solving PS, Metacognition MC, Self-control SC, Empathy E, Cooperation CO, Emotional intelligence EI, Communication Com, Teamwork TW, Adaptability A, Leadership L, Digital literacy DL, Creativity CR, Critical thinking CT, Collaboration CL, Cultural awareness CA, Innovation I. This targeted selection allows us to demonstrate the feasibility and consistency of the proposed methodological workflow within a controlled framework. The construction of the contingency matrix is based on a semi-structured annotation process that combines an initial phase of automatic keyword extraction and automatic calculation of keyword occurrence frequencies within the selected data sources, followed by validation by three experts who narrow the analysis down to the relevant keywords identified and carry out annotation in accordance with specific guidelines. Likert-scale scores from 0 to 5 (0–Not relevant at all, 1–Almost relevant, 2–Not very relevant, 3–Moderately relevant, 4–Fairly relevant, 5–Highly relevant) reflect both frequency and contextual relevance of terms within each source (see Table 2). To ensure consistency and limit subjectivity, a unified annotation guideline was applied throughout the process. Scoring rules were defined prior to analysis, and ambiguous cases were resolved through cross-validation against predefined criteria. The guideline clearly defines the verbal anchors and keyword’s thematic relevance within a data source for each Likert scale value from 0 to 5:

- 5—Highly relevant: a central, recurring keyword that is essential for summarising the main theme of the data source.
- 4—Fairly relevant: an important keyword, closely related to the topic, but not entirely central.
- 3—Moderately relevant: a keyword related to the topic, but not essential; could be removed without losing the main point.
- 2—Not very relevant: a peripheral keyword, mentioned once or twice, with no strong link to the core of the data source.
- 1—Almost relevant: a marginal keyword; the annotator is unsure, but perceives a very weak link.
- 0—Not relevant at all: keyword not related to the topic, off-topic, or clearly irrelevant here.

Table 2. Contingency matrix ensuring correspondence between data sources and relevant keywords relating to the semantic class Meta Skills.

S	K	M	PS	MC	SC	E	CO	EI	Com	TW	A	L	DL	CR	CT	CL	CA	I
S1	0	0	0	0	3	4	3	5	0	0	0	0	0	0	0	0	0	0
S2	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	5	3	3
S3	5	2	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S4	0	0	0	0	0	0	0	0	4	3	3	1	3	0	0	0	0	0
S5	0	0	0	0	4	3	2	5	0	0	0	0	0	0	0	0	0	0
S6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	4	3
S7	4	5	5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S8	0	0	0	0	0	0	0	0	4	4	3	2	2	0	0	0	0	0
S9	0	0	0	0	2	2	4	5	0	0	0	0	0	0	0	0	0	0
S10	0	0	0	0	0	0	0	0	2	2	4	2	2	0	0	0	0	0
S11	0	0	0	0	2	3	1	4	1	0	0	0	0	0	0	0	0	0
S12	2	5	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S13	0	0	0	0	0	0	0	0	4	3	2	2	4	0	0	0	0	0
S14	0	0	0	0	5	4	4	5	0	0	0	0	0	0	0	0	0	0
S15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	4	2
S16	3	1	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S17	0	0	0	0	0	0	0	0	4	4	4	2	4	0	0	0	0	0
S18	0	0	0	0	0	0	0	0	2	1	4	2	3	0	0	0	0	0

The annotation process involved three domain experts. Although a formal inter-annotator agreement coefficient (e.g., Cohen's or Fleiss' kappa) was not computed, consistency was ensured through an iterative consensus-based validation process. Discrepancies were discussed and resolved collectively. This represents a limitation and will be addressed in future work through formal agreement metrics. The annotation procedure was designed to ensure coherence across expert judgments while preserving the exploratory nature of the analysis.

IBM SPSS Statistics 23 is the software chosen to perform the simulation with the following parameters (Method: Correlation matrix; Extraction Method: Principal Component Analysis; Rotation Method: Varimax with Kaiser Normalization; number of extracted factors: 2; maximum number of iterations: 25). Principal Component Analysis (PCA) is used in this study as an exploratory descriptive technique, without any claim of construct validation or latent structure confirmation.

The projection of points-columns onto the first factorial plane formed by the first two eigenvectors (see Figure 9) corresponding to the first two retained components of the dual analysis of the correspondence matrix between data sources and relevant keywords makes it possible to detect four sub-classes of the main semantic class *Meta Skills: Cognitive Skills* (Knowledge K, Memory M, Problem-solving PS and Metacognition MC), *Social Skills* (Self-control SC, Empathy E, Cooperation CO and Emotional intelligence EI), *Professional Skills* (Communication Com, Teamwork TW, Adaptability A, Leadership L and Digital literacy DL) and *Cultural Skills* (Creativity CR, Critical thinking CT, Collaboration CL, Cultural awareness CA and Innovation I) (see Table 3).

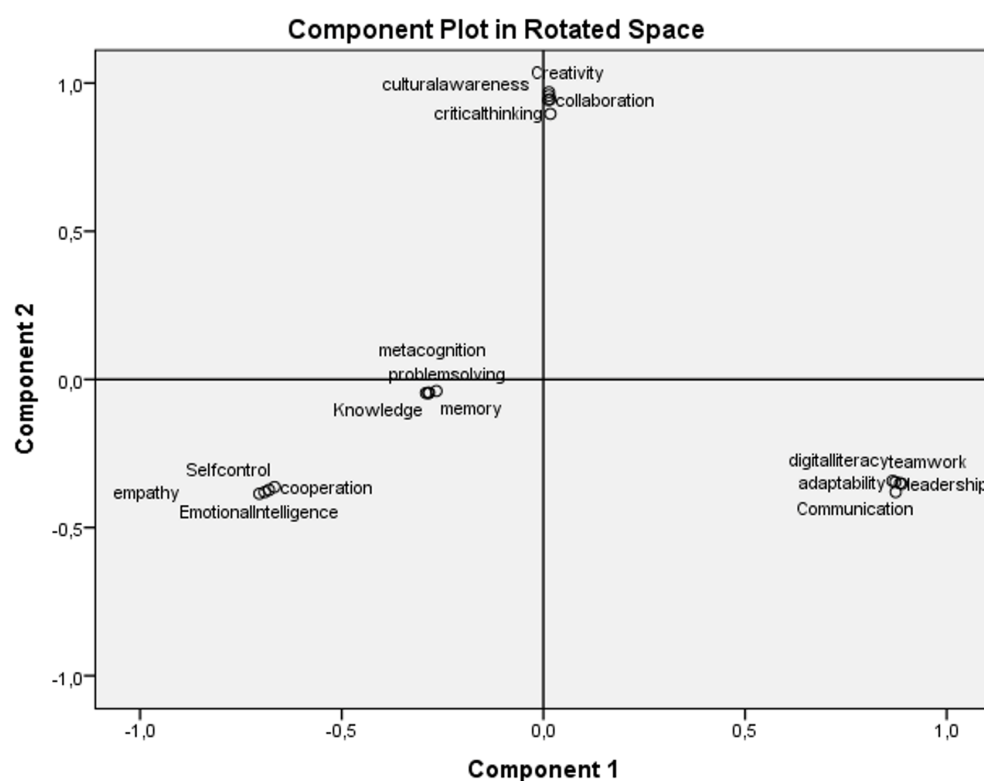


Figure 9. Projection of points-columns onto the first factorial plane.

The projection of points-rows onto the first factorial plane formed by the first two eigenvectors corresponding to the first two retained components of the direct analysis of the correspondence matrix between data sources and relevant keywords makes it possible to detect the presence of semantic clusters. Each cluster, characterized by strong internal cohesion and weak external coupling, contains data sources that exhibit a high degree of similarity and are close in terms of chi-squared distance (χ^2)

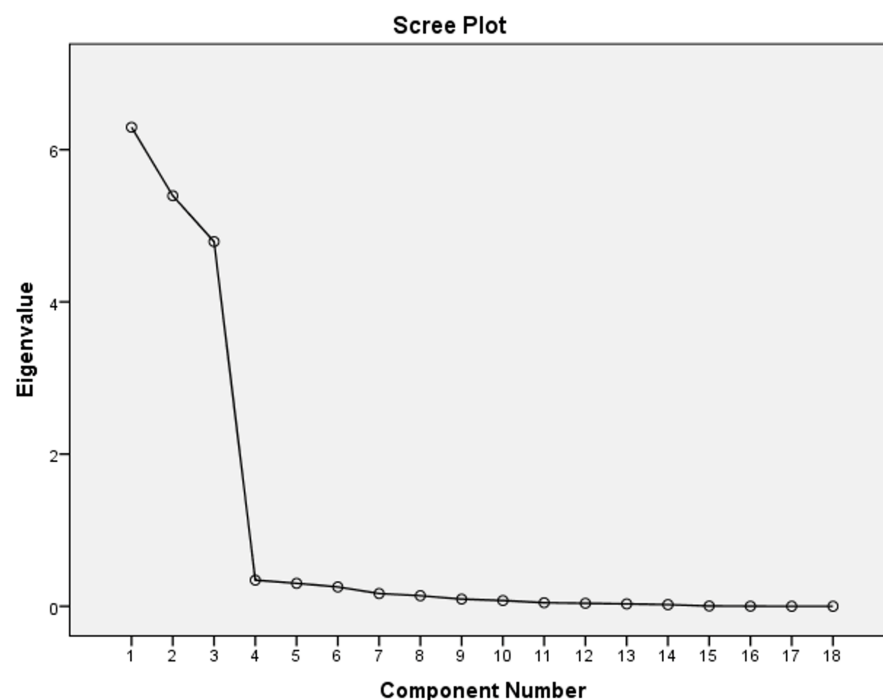
Table 3. Subdivision of the semantic class *Meta Skills* into four subclasses.

Semantic Subclasses	Relevant Keywords
Cognitive Skills	Knowledge K, Memory M, Problem-solving PS, Metacognition MC
Social Skills	Self-control SC, Empathy E, Cooperation CO, Emotional intelligence EI
Professional Skills	Communication Com, Teamwork TW, Adaptability A, Leadership L, Digital literacy DL
Cultural Skills	Creativity CR, Critical thinking CT, Collaboration CL, Cultural awareness CA, Innovation I

The main advantage of factor analysis is that it offers the possibility of superimposing the two projections relating to scatter plot of points-rows $N(I)$ and scatter plot of points-columns $N(J)$ in a lower dimension space, in our case the first factorial plane of dimension 2. The dual scatter plot $N(J) = \{y_j \in R^p, j \in J\}$ corresponding to points-columns is constituted by $Y_j = (y_{1j}, y_{2j}, \dots, y_{nj})$ with $y_{ij} = f_{ij}/f_{.j}$ and has vector $H = (f_{1.}, f_{2.}, \dots, f_{n.})$ as its barycenter.

In direct analysis operating on the scatter plot $N(I)$, the first eigenvector (axis of maximum elongation of the scatter plot) explains most of the total inertia. The axes of inertia in R^n are indexed by $\alpha = 1, 2, \dots, n$. It is important to note that the inertia λ_α explained by the α -axis is identical in direct analysis operating on data sources and dual analysis operating on keywords. In addition, the eigenvectors are orthogonal in space, whether in direct or dual analysis; this means that the cumulative percentage of total inertia explained by the first two eigenvectors in both analyses is equal to $\frac{\lambda_1 + \lambda_2}{\sum_{\alpha=1}^n \lambda_\alpha}$.

The quality of the analysis can be assessed using the eigenvalue histogram. This quality depends on the percentage of inertia explained by the first two eigenvectors $\lambda_{\alpha_1} + \lambda_{\alpha_2}$. A factor analysis is generally considered acceptable when the selected factors explain the majority of the total variance. The first two retained components explain 64,939 per cent of the total variance. This level is interpreted as moderate in the context of exploratory PCA and should not be considered a benchmark of model adequacy (see Figure 10).

**Figure 10.** Histogram of eigenvalues.

Cronbach's alpha measures the internal consistency of a cluster comprising a group of items intended to reflect the same latent construct. An alpha value close to 1 indicates that the items within the cluster are consistent with one another. A high alpha value suggests that the cluster in question is relatively well-defined and can be treated as a reliable scale (see Table 4).

Table 4. Reliability statistics of the generated clusters.

Cluster	Semantic Subclasses	Number of Items	Cronbach's Alpha
C1	Cognitive Skills	4	0.962
C2	Social Skills	4	0.969
C3	Professional Skills	5	0.967
C4	Cultural Skills	5	0.971

The obtained Cronbach's alpha values (>0.96) provide preliminary evidence of internal consistency and cluster coherence, although they should not be interpreted as definitive psychometric validation given the limited number of items per group and the diagonal block structure of the contingency matrix.

We denote $F_{\alpha}(i)$ coordinate of X_i on the axis α of R^p and $G_{\alpha}(j)$ coordinate of Y_j on the axis α of R^n . There are Relationship transitions between $F_{\alpha}(i)$ and $G_{\alpha}(j)$:

$$F_{\alpha}(i) = \frac{1}{\sqrt{\lambda_{\alpha}}} \sum_{j=1}^p \frac{f_{ij}}{f_{i.}} G_{\alpha}(j) \quad (4)$$

$$G_{\alpha}(j) = \frac{1}{\sqrt{\lambda_{\alpha}}} \sum_{i=1}^n \frac{f_{ij}}{f_{.j}} F_{\alpha}(i) \quad (5)$$

The objective of simultaneous projections of scatter plot of data sources $N(I)$ and dual scatter plot of keywords $N(J)$ on the first factorial plane is to highlight latent correspondences, detect affinities between vectors and identify the hidden variables responsible for some affinities. The proximity between two points-rows (such as data sources S3, S7, S12 and S16) reflects a similar keywords profile. On the other hand, the proximity between two points-columns (such as keywords: Knowledge K, Memory M, Problem-solving PS and Metacognition MC) reflects a similar profile of data sources. This can be exploited to perform an efficient search for relevant data sources close to the keywords used in this search. We can note that the data sources S3, S7, S12 and S16 rather cover the area of *Cognitive Skills* (see Figure 11) while the data sources S1, S5, S9, S11 and S14 tend to cover the area of *Social Skills*. In addition, the data sources S4, S8, S10, S13, S17 and S18 cover the area of *Professional Skills* while the data sources S2, S6 and S15 focus on the field of *Cultural Skills* (see Table 5).

Table 5. Generated clusters.

Cluster	Semantic Subclasses	Relevant Data Sources
C1	Cognitive Skills	S3, S7, S12, S16
C2	Social Skills	S1, S5, S9, S11, S14
C3	Professional Skills	S4, S8, S10, S13, S17, S18
C4	Cultural Skills	S2, S6, S15

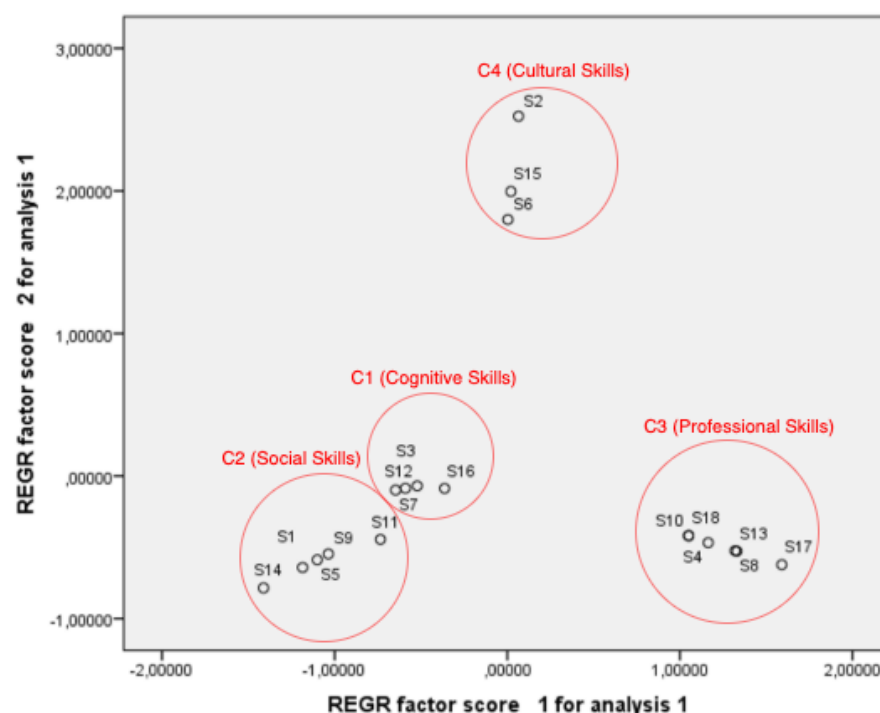


Figure 11. Projection of points-rows onto the first factorial plane.

We chose to subdivide each semantic class into smaller, more specialized subclasses to reduce the complexity of semantic indexing and ensure better scalability while avoiding bottlenecks. This process is accompanied by an increase in the degree of the Hilbert curve as the ontology evolves (see Figure 12). This operation is not performed dynamically within the peer-to-peer network. It is carried out offline during a scheduled semantic boot phase during a period of minimal load. It should be noted that the Hilbert curve tends to fill the unit square when n , the number of iterations, becomes very large. This means that each peer in our system will be responsible for one or more concepts, and will therefore have a part of the overall ontology to manage, and must consequently respond to the various requests involving the concepts within its competence. This requirement is consistent with the principles of peer-to-peer architectures, which stipulate that each node can act as both a client and a server.

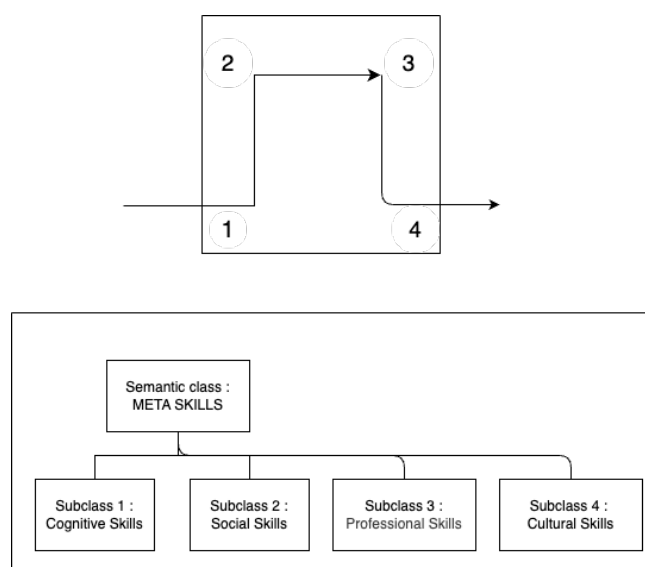


Figure 12. Evolution of Hilbert curve with semantic refinement.

5.3. Semantic Indexing of a New Data Source

The insertion of new data sources will be carried out as follows: The data source to be inserted propagates from one semantic class to another via an epidemic protocol [16]. Any new data source to be inserted will be considered a fictional information source within each semantic class (see Table 6).

Table 6. Inserting a new data source into the contingency matrix.

	T_1	T_2	...	T_p
S_1	f_{11}	f_{12}	...	f_{1p}
S_2	f_{21}	f_{22}	...	f_{2p}
.	.	.	...	.
.	.	.	...	.
S_n	f_{n1}	f_{n2}	...	f_{np}
$S_{fictional}$	f_{n+11}	f_{n+12}	...	f_{n+1p}

We know that the variables that are best represented on the factorial plane are those with the strongest coordinates. This means that, for example, the weak contribution of the fictitious data source to forming the factors (representing the underlying dimensions which, once interpreted, describe the data in a synthetic way) leads us to refute the hypothesis that it belongs to this semantic class (see Figure 13).

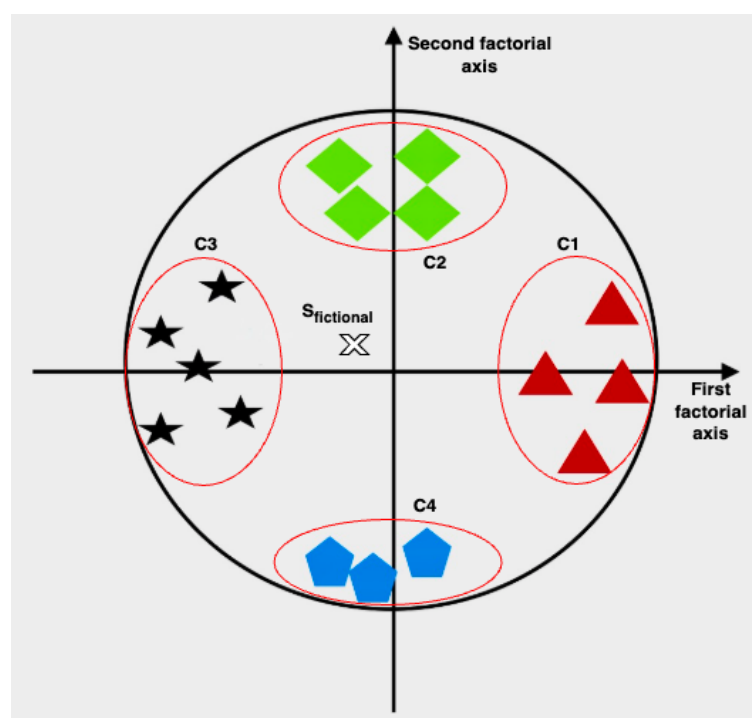


Figure 13. Projection of scatter plot onto the first factorial plane.

The data source to be inserted can belong to one or more semantic classes. For example, a data source from the field of biochemistry can belong to the semantic classes of chemistry and biology. This will allow our proposal to manage the overlaps between the different domains. The $\cos^2$ projection, or cosine squared ($\cos^2$), measures the quality of representation of a point (individual, variable, or category) on a factorial axis or plane in factor analysis. It indicates the proportion of variance or inertia explained by this projection. The $\cos^2$ is the square of the cosine of the angle formed between the vector of the point in

the initial space and its projection onto the factorial axis. Its value ranges from 0 to 1: a $\cos^2$ close to 1 indicates an excellent representation, while a low $\cos^2$ (close to 0) indicates a poor projection. To this end, it can be stated that a semantic class is considered to index $S_{fictional}$ if its squared cosine is strictly greater than a relevance threshold that can be configured within the system. Thus, several semantic classes may index $S_{fictional}$ provided that the corresponding squared cosine is greater than the relevance threshold.

A threshold of 0.5 for cosine squared values was adopted as a heuristic criterion to indicate moderate-to-strong association between items and components. This threshold is not theoretically fixed and is used here for exploratory interpretation purposes. Preliminary analyses using different thresholds (0.4 and 0.6) indicated that lower thresholds increased class overlap, while higher thresholds improved class separation at the expense of reduced coverage. The chosen threshold represents a compromise between stability and inclusivity.

5.4. Storage of Resources Related to the Semantic Class

We propose to combine horizontal partitioning and vertical partitioning mechanisms to ensure the storage of the contingency matrix as well as the table containing the addresses of the peers that physically store resources indexed at the semantic class level. This means that each node member of the semantic class stores a fragment resulting from a joint horizontal and vertical partitioning, this satisfies P2P paradigm philosophy which stipulates that each peer can play the role of both client and server. We remind that the different member nodes of a semantic class are linked together by means of the hypercube structure. This provides effective technical support for information dissemination between various member nodes. In the following example (see Figure 14), we restrict ourselves, without loss of generality, to a contingency matrix consisting of 5 resources.

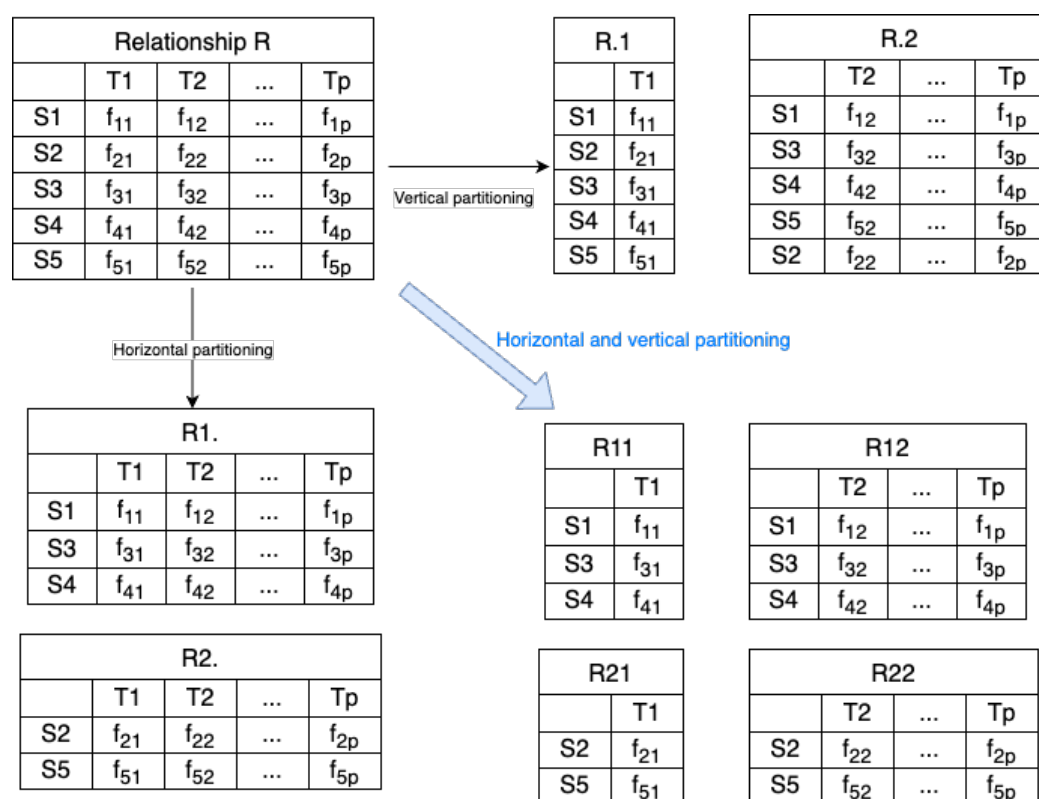


Figure 14. Partitioning the contingency matrix.

5.5. Search for Relevant Data Sources

In structured P2P systems (Chord, CAN, Tapestry, Pastry, etc.), peers and resources belong to the same namespace and are characterized by identifiers considered unique within that same space. Each identifier is calculated by applying a hash function. The keys corresponding to resources are often placed under responsibility of peer with the closest numerical identifier. Each peer has a routing table that directs messages to a peer closer to their destination. Certainly, structured P2P mode offers the possibility of simple, efficient and reliable resource searching (e.g., any data can be accessed in $O(\log(n))$ with an underlying Chord network with n participants), and presents good performance in terms of scalability and fault tolerance. In contrast, structured P2P systems based on a DHT (Distributed Hash Table) are suitable for unidimensional searches involving a single search attribute. In addition, the exact matching of keys for access to resources is a major obstacle to the structured DHT P2P system: we need to know the exact key identifying a resource to locate the peer responsible for its storage. This problem is exacerbated in the case of DHT since most hash functions produce radically different outputs if the input is slightly modified. However, in reality users often only have partial information to identify these resources, and tend to submit broadly intentional requests. In addition, the P2P DHT mode suffers from an intrinsic semantic problem. Indeed, applying the hash function to two semantically related pieces of information will, in principle, produce two independent keys, which will then be stored in two independent pairs.

In this article, the proposed approach designs algorithms to overcome the problem of exact matching search and the limitations of one-dimensional search based on a single search attribute. The example shown in Figure 15 illustrates the processing flow for a broad query. Meanwhile, the example in Figure 16 presents the processing flow for multidimensional queries involving multiple attributes

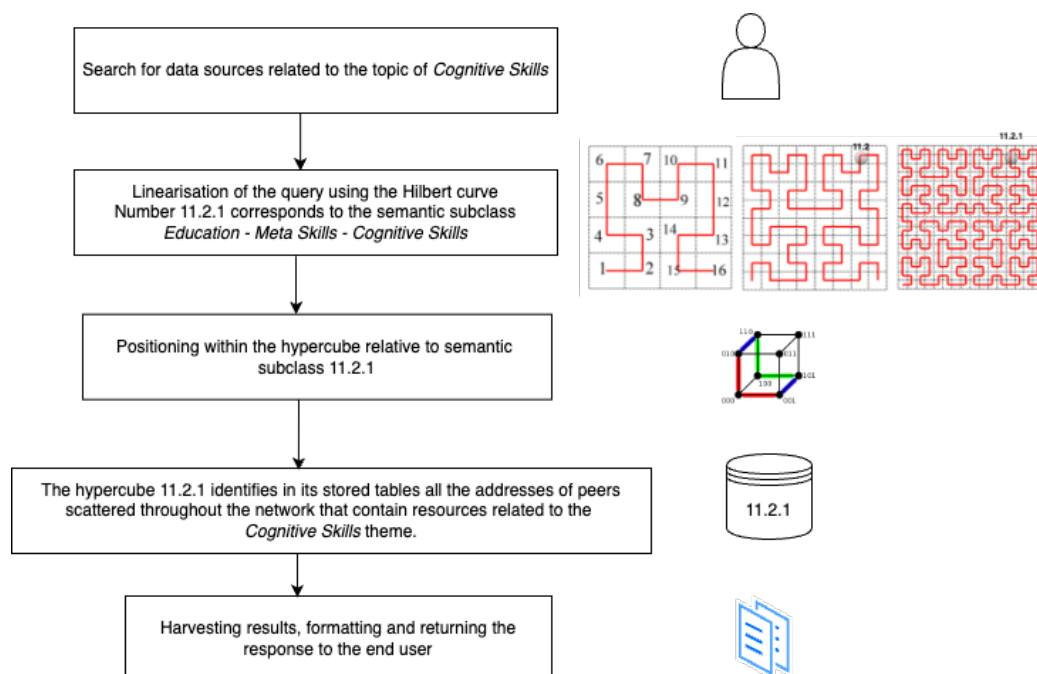


Figure 15. Processing flow for a broad query.

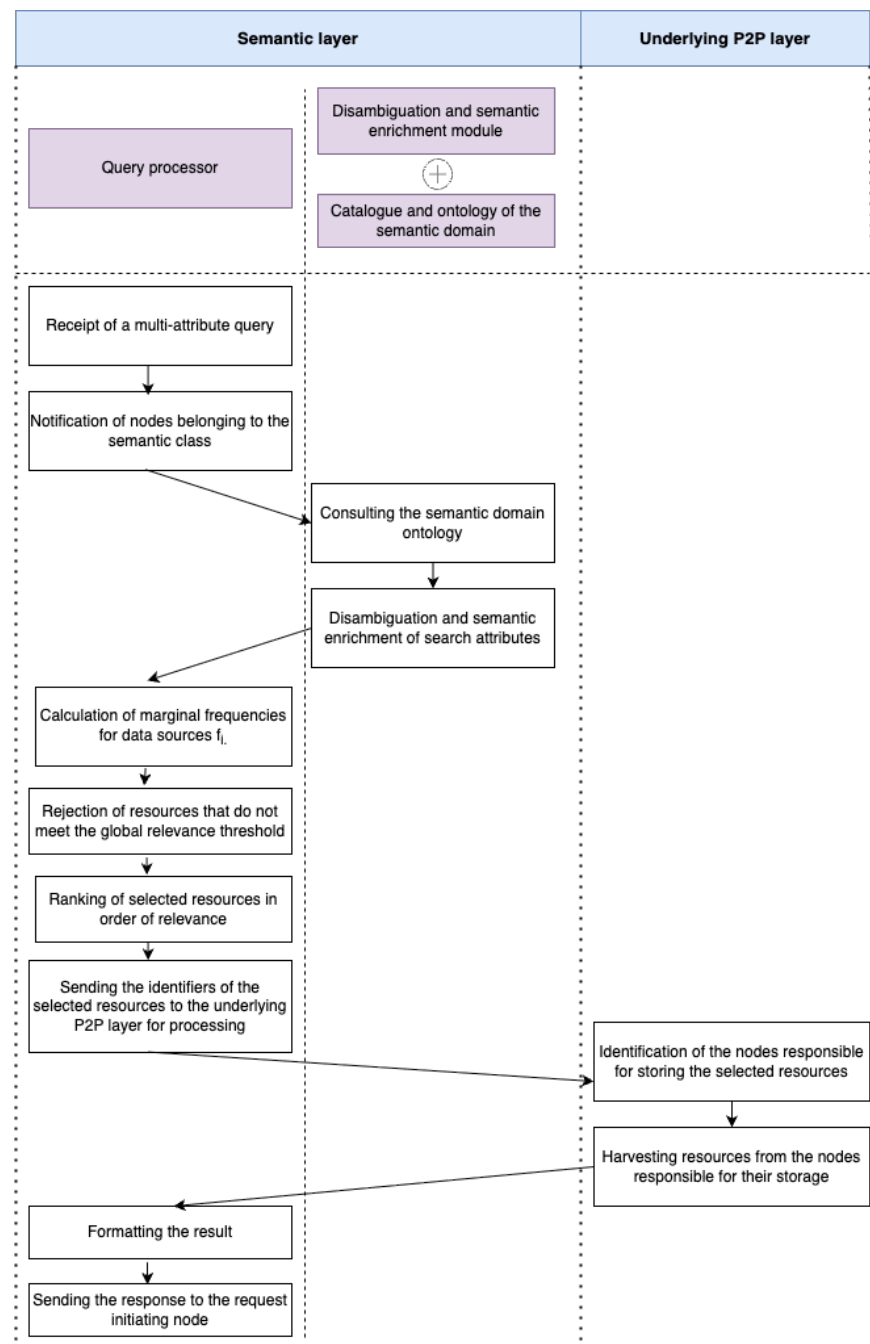


Figure 16. Processing multi-attribute queries.

6. Discussion

We outline our contributions in the following points:

- Contribution 1: Indexing of semantic classes using the order induced by Hilbert curve, renowned for its valuable characteristics in preserving locality and semantic affinity. This curve offers better preservation of locality during scaling, notably by supporting semantic refinement induced by ontology evolution. This improves the preservation of semantic affinity during scaling.
- Contribution 2: Organizing peer members of a semantic class using a hypercube architecture. This graph, renowned for its efficient dissemination algorithm, keeps the semantic class structure up-to-date while significantly reducing the number of messages exchanged between peers. The hypercube offers a structure that improves traffic optimization as well as fault tolerance.

- Contribution 3: Clustering of data sources based on keyword semantics. This operation is performed using factor analysis which reveals latent semantic similarities between different resources. Furthermore, we propose a method for constructing the contingency matrix based on calculating occurrence frequency of keywords within a resource.
- Contribution 4: The proposed semantic layer designs algorithms to overcome the problem of exact matching search and the limitations of one-dimensional search based on a single search criterion.

However, there are a number of limitations to our work:

- The results are based on expert-driven annotation and exploratory statistical analysis, which may introduce subjectivity and limit generalizability.
- A limitation of this study lies in the absence of a formal inter-annotator agreement metric (e.g., Cohen's kappa or Krippendorff's alpha). Although annotation guidelines were applied, the reliability of expert judgments remains partially unquantified, which is consistent with the exploratory nature of this work.
- No direct comparison with standard text clustering or classification methods (e.g., LDA, TF-IDF with k-means, or embedding-based approaches) was conducted in this study. The focus remains on exploratory structuring rather than performance benchmarking.

Several avenues for improving our work can be considered.

- Implement our proposal with several semantic classes containing a very large number of peers and data sources.
- Perform a network simulation (churn, latency, routing hops)
- Several approaches are being explored to address the issue of dummy nodes, in particular logical node aggregation, virtual peer mapping, and probabilistic neighbor reuse, which can mitigate dummy-node overhead

7. Conclusions

Semantic indexing of data sources is of paramount importance for transforming masses of raw data into structured, interoperable and usable knowledge. Several approaches to semantic indexing are available. However, the main challenge remains scaling up. This article focuses on a conceptual framework designed for scalable semantic indexing of data sources based on structured peer-to-peer system adapted for managing very large number of nodes, Hilbert curve renowned for its preservation of semantic affinity while scaling, hypercube structure with its efficient diffusion algorithm, semantic clustering of data sources based on keywords, as well as machine learning techniques, in particular, multidimensional data analysis. In addition, the proposed approach designs algorithms to overcome the problem of exact matching search and the limitations of one-dimensional search based on a single search attribute. This study provides an exploratory conceptual framework rather than a validated or operational system. Future work will focus on empirical validation, scalability testing across multiple semantic classes using numerous peers and data sources, and comparison with established computational approaches. We also plan to improve our approach by incorporating AI techniques such as NLP, entity recognition and RAG.

Author Contributions: Conceptualization, M.A., F.A. and A.B.; Methodology, M.A., F.A. and A.B.; Software, M.A., F.A. and A.B.; Validation, M.A., F.A. and A.B.; Formal analysis, M.A., F.A. and A.B.; Investigation, M.A., F.A. and A.B.; Resources, M.A., F.A. and A.B.; Data curation, M.A., F.A. and A.B.; Writing—original draft, M.A., F.A. and A.B.; Writing—review & editing, M.A., F.A. and A.B.; Visualization, M.A., F.A. and A.B.; Supervision, M.A., F.A. and A.B.; Project administration, M.A., F.A. and A.B.; Funding acquisition, M.A., F.A. and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analysed during this study are included in this published article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

P2P	Peer-to-Peer
K	Knowledge
M	Memory
PS	Problem-solving
MC	Metacognition
SC	Self-control
E	Empathy
CO	Cooperation
EI	Emotional intelligence
Com	Communication
TW	Teamwork
A	Adaptability
L	Leadership
DL	Digital literacy
CR	Creativity
CT	Critical thinking
CL	Collaboration
CA	Cultural awareness
I	Innovation

Appendix A

Table A1. Data sources related to the semantic class Meta Skills.

Source of Information	Reference
S1	[17]
S2	[18]
S3	[19]
S4	[20]
S5	[21]
S6	[22]
S7	[23]
S8	[24]
S9	[25]
S10	[26]
S11	[27]
S12	[28]
S13	[29]
S14	[30]
S15	[31]
S16	[32]
S17	[33]
S18	[34]

References

1. Thirumahal, R.; Sadasivam, G.S.; Shruti, P. Semantic Integration of Heterogeneous Data Sources Using Ontology-Based Domain Knowledge Modeling for Early Detection of COVID-19. *SN Comput. Sci.* **2022**, *3*, 428. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Fusco, G.; Aversano, L. An Approach for Semantic Integration of Heterogeneous Data Sources. *PeerJ Comput. Sci.* **2020**, *6*, e254. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Sun, Y.; Chun, S.-J.; Lee, Y. Learned Semantic Index Structure Using Knowledge Graph Embedding and Density-Based Spatial Clustering Techniques. *Appl. Sci.* **2022**, *12*, 6713. [\[CrossRef\]](#)
4. Haghgoo, M.; Nazary Aghche Mazary, A.; Monti, A. SiSEG-Auto Semantic Annotation Service to Integrate Smart Energy Data. *Energies* **2022**, *15*, 1428. [\[CrossRef\]](#)
5. Chong, S.S.; Schildhauer, M.; O'Brien, M.; Mecum, B.; Jones, M.B. Enhancing the FAIRness of Arctic Research Data Through Semantic Annotation. *Data Sci. J.* **2024**, *23*, 1–14. [\[CrossRef\]](#)
6. Stoica, I.; Morris, R.; Liben-Nowell, D.; Karger, D.R.; Kaashoek, M.F.; Dabek, F.; Balakrishnan, H. Chord: A Scalable Peer-to-Peer Lookup Protocol for Internet Applications. *IEEE/ACM Trans. Netw.* **2003**, *11*, 17–32. [\[CrossRef\]](#)
7. Saravanan, S.; Ashfauk Ahamed, A. AI-Driven Web Crawling for Semantic Extraction of News Content from Newspapers. *Sci. Rep.* **2025**, *15*, 41673. [\[CrossRef\]](#)
8. Raj, Y.; Mahdi, H.M.; Abraham, B.J.; Rama Sree, S.; Kiruthika, R.; Ugli, K.I.J. Real-Time Semantic Indexing for High-Volume Data Streams. *Indian J. Inf. Sources Serv.* **2025**, *15*, 423–431. [\[CrossRef\]](#)
9. Hong, K.; Park, Y. Large Language Models for Semantic Join: A Comprehensive Survey. *IEEE Access* **2025**, *13*, 184478–184493. [\[CrossRef\]](#)
10. Kang, S.; Zhang, Y.; Jiang, P.; Lee, D.; Han, J.; Yu, H. Taxonomy-Guided Semantic Indexing for Academic Paper Search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 7169–7184. [\[CrossRef\]](#)
11. Slavic, A. Use of the Universal Decimal Classification: A World-Wide Survey. *J. Doc.* **2008**, *64*, 211–228. [\[CrossRef\]](#)
12. Yang, Y.; Zuo, X.; Zhao, K.; Li, Y. Integrating NoSQL, Hilbert Curve, and R*-Tree to Efficiently Manage Mobile LiDAR Point Cloud Data. *Isprs Int. J. Geo-Inf.* **2024**, *13*, 253. [\[CrossRef\]](#)
13. Prusinkiewicz, P.; Lindenmayer, A. *The Algorithmic Beauty of Plants*; Springer: New York, NY, USA, 1990.
14. Cicerone, S.; Di Fonso, A.; Di Stefano, G.; Navarra, A.; Piselli, F. Mutual and Total Mutual Visibility in Hypercube-like Graphs. *Appl. Math. Comput.* **2025**, *491*, 129216. [\[CrossRef\]](#)
15. Schlosser, M.; Sintek, M.; Decker, S.; Nejdl, W. HyperCuP—Hypercubes, Ontologies, and Efficient Search on Peer-to-Peer Networks. *Lect. Notes Comput. Sci.* **2002**, *2530*, 112–124.
16. Shashidhar, N.; Kari, C.; Verma, R. The Efficacy of Epidemic Algorithms on Detecting Node Replicas in Wireless Sensor Networks. *J. Sens. Actuator Netw.* **2015**, *4*, 378–409. [\[CrossRef\]](#)
17. Belfi, B.; Borghans, L. The Importance of Socio-Emotional Skills for Multiple Life Outcomes and the Role of Education. 2025. Available online: <https://ssrn.com/abstract=5086577> (accessed on 12 April 2026).
18. Thornhill-Miller, B.; Camarda, A.; Mercier, M.; Burkhardt, J.; Morisseau, T.; Bourgeois-Bougrine, S.; Vinchon, F.; Hayek, S.; Augereau-Landais, M.; Mourey, F.; et al. Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. *J. Intell.* **2023**, *11*, 54. [\[CrossRef\]](#)
19. Royer, J.; Cisero, C.; Carlo, M. Techniques and Procedures for Assessing Cognitive Skills. *Rev. Educ. Res.* **1993**, *63*, 201–243. [\[CrossRef\]](#)
20. Tushar, H.; Sooraksa, N. Global Employability Skills in the 21st Century Workplace: A Semi-Systematic Literature Review. *Heliyon* **2023**, *9*, e21023. [\[CrossRef\]](#)
21. Guo, J.; Tang, X.; Marsh, H.; Parker, P.; Basarkod, G.; Sahdra, B.; Ranta, M.; Salmela-Aro, K. The Roles of Socio-Emotional Skills in Students' Academic and Life Success: A Multi-Informant and Multicohort Perspective. *J. Pers. Soc. Psychol.* **2021**, *124*, 1079. [\[CrossRef\]](#)
22. Bridgstock, R.; Cunningham, S. Creative Labour and Graduate Outcomes: Implications for Higher Education and Cultural Policy. *Int. J. Cult. Policy* **2016**, *22*, 10–26. [\[CrossRef\]](#)
23. Anderson, J. Acquisition of Cognitive Skill. *Psychol. Rev.* **1982**, *89*, 369–406. [\[CrossRef\]](#)
24. Suarta, M.; Suwintana, I.; Sudhana, I.; Hariyanti, N. Employability Skills Required by the 21st Century Workplace: A Literature Review of Labor Market Demand. In *International Conference on Technology and Vocational Teachers (ICTVT 2017)*; Atlantis Press: Dordrecht, The Netherlands, 2017.
25. Wang, F.; King, R.; Zeng, L. Cooperative School Climates Are Positively Linked with Socio-Emotional Skills: A Cross-National Study. *Br. J. Educ. Psychol.* **2024**, *94*, 622–641. [\[CrossRef\]](#)
26. Tomasson Goodwin, J.; Goh, J.; Verkoeyen, S.; Lithgow, K. Can Students Be Taught to Articulate Employability Skills? *Educ. Train.* **2019**, *61*, 445–460. [\[CrossRef\]](#)

27. Kim, E.; Allen, J.; Jimerson, S. Supporting Student Social Emotional Learning and Development. *Sch. Psychol. Rev.* **2024**, *53*, 201–207. [[CrossRef](#)]
28. Lopez, J.; Behrman, J.; Cueto, S.; Favara, M.; Sánchez, A. Late-Childhood Foundational Cognitive Skills Predict Educational Outcomes through Adolescence and into Young Adulthood: Evidence from Ethiopia and Peru. *Econ. Educ. Rev.* **2024**, *100*, 102543. [[CrossRef](#)]
29. McGunagle, D.; Zizka, L. Employability Skills for 21st-Century STEM Students: The Employers' Perspective. *High. Educ. Skills Work.-Based Learn.* **2020**, *10*, 591–606. [[CrossRef](#)]
30. Eisner, M.; Ribeaud, D.; Sorrenti, G.; Zölitz, U. The Causal Impact of Socio-Emotional Skills Training on Educational Success. 2020. Available online: <https://ssrn.com/abstract=3560331> (accessed on 12 April 2026).
31. Collard, P.; Looney, J. Nurturing Creativity in Education. *Eur. J. Educ.* **2014**, *49*, 348–364. [[CrossRef](#)]
32. Kerr, S.; Pekkarinen, T.; Uusitalo, R. School Tracking and Development of Cognitive Skills. *J. Labor Econ.* **2013**, *31*, 577–602. [[CrossRef](#)]
33. Dwidienawati, D.; Istiani, C.; Rahman, R. New Entrants Joining the Industry, What Skill Do They Need? *J. Infrastruct. Policy Dev.* **2024**, *8*, 8546. [[CrossRef](#)]
34. Zheng, X.; Mohd Puad, M.H.; Ab Jalil, H.; Zhendong, G.; Wei, W. Reconceptualization and Pedagogical Strategies of Public Speaking Competency for Employability: An Integrative Review. *Educ. Train.* **2025**, *67*, 56–88. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.