2022 IEEE International Conference on Knowledge Graph (ICKG) | 978-1-6654-5101-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICKG55886.2022.00023

An efficient storage scheme for Sustainable Development Goals data over distributed knowledge graph stores

Irene Kilanioti (a.k.a. Eirini Koilanioti)
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
eirinikoilanioti@mail.ntua.gr

George Angelos Papadopoulos
Department of Computer Science
University of Cyprus
Aglantzia, Cyprus
george@cs.ucy.ac.cy

***Abstract*—The achievement of the Sustainable Development Goals (SDGs) is important in order to ensure a world worth living in for future generations. Digitization and the plethora of data available for analysis offer new opportunities to support and monitor the achievement of the SDGs. Scholars can contribute to the achievement of the SDGs by guiding the actions of practitioners based on the analysis of data, as intended by this work. In this paper, we propose dimensionality reduction methods to semantically cluster new uncategorised SDG data and novel indicators, and efficiently place them in the environment of a distributed knowledge graph store. In particular, our work proposes and experimentally corroborates the use of Hilbert Space Filling Curves (HSFCs) to efficiently store real SDG data with reduced retrieval times and preservation of their semantic closeness. First, algorithm is theoretically founded and explained and an approach for data classification of entrant-indicators is described. Then, a thorough case study in a distributed knowledge graph environment experimentally evaluates our algorithm. The results are presented and discussed in light of theory along with the actual impact that can have for practitioners analysing SDG data, including intergovernmental organizations, government agencies and social welfare organizations. Our approach empowers SDG knowledge graphs for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions, as data are accessed in reduced retrieval times. It facilitates quicker measurement of influence of users and communities on specific goals and serves for faster distributed knowledge matching, as semantic cohesion of data is preserved.**

***Index Terms*—Sustainable Development Goals ontology, distributed knowledge graphs, Hilbert Space Filling Curves, distributed storage, sustainability**

I. Introduction

The collective effort to optimally harmonize sustainability goals bears a transformative view of the world and requires the conscious social, fiscal and technological contribution of many societal agents among which sustainable IT can also play a crucial role. “Data which are high-quality, accessible, timely, reliable and disaggregated by characteristics relevant in national contexts” is required (A/RES/70/01) [1].

In the framework of the United Nations (UN) 2030 Agenda for Sustainable Development a measurable international initiative to eradicate poverty, safeguard the environment and maintain peace and welfare was established: the Sustainable Development Goals (SDGs) [2]. SDG ontology comprises substantially sustainable development goals, targets, indicators and data series for the quantification of their accomplishment (Fig. 1) [3], and the full taxonomy is accessible as linked open data at [4]. Depending on the grade of development of internationally established methodology and standards as well as regularity of production of relevant data, the aforementioned indicators are categorised into tiers.

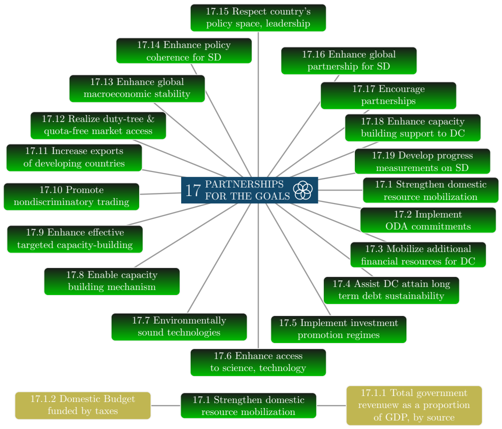

17.15 Respect country’s policy space, leadership
17.14 Enhance policy coherence for SD
17.16 Enhance global partnership for SD
17.13 Enhance global macroeconomic stability
17.17 Encourage partnerships
17.12 Realize duty-tree & quota-free market access
17.18 Enhance capacity building support to DC
17.11 Increase exports of developing countries
17.19 Develop progress measurements on SD
17 PARTNERSHIPS FOR THE GOALS
17.10 Promote nondiscriminatory trading
17.1 Strengthen domestic resource mobilization
17.2 Implement ODA commitments
17.9 Enhance effective targeted capacity-building
17.3 Mobilize additional financial resources for DC
17.8 Enable capacity building mechanism
17.4 Assist DC attain long term debt sustainability
17.7 Environmentally sound technologies
17.5 Implement investment promotion regimes
17.6 Enhance access to science, technology
17.1.2 Domestic Budget funded by taxes
17.1 Strengthen domestic resource mobilization
17.1.1 Total government revenuew as a proportion of GDP, by source

Fig. 1. SDG Goal 17 consists of targets, such as Target 17.1: ”Strengthen domestic resource mobilization” and entails indicators such as 17.1.2: ”Total government revenue as a proportion of GDP, by source” [2].

Indeed, the question of how we can leverage data to estimate the impact of various actions especially in the context of social welfare and sustainability is a highly relevant topic in IT research and of great interest for society. Hence, we need to reduce access times for SDG data analysis and improve semantic cohesion of uncategorized data. Efficient processing and storage solutions for data in this respective field are necessary for practitioners, that entail intergovernmental organizations, government agencies and social welfare organizations, i.e. civic organizations and associations of persons engaged in the promotion of social welfare.

In this work, we propose dimensionality reduction methods

978-1-6654-5101-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ICKG55886.2022.00023

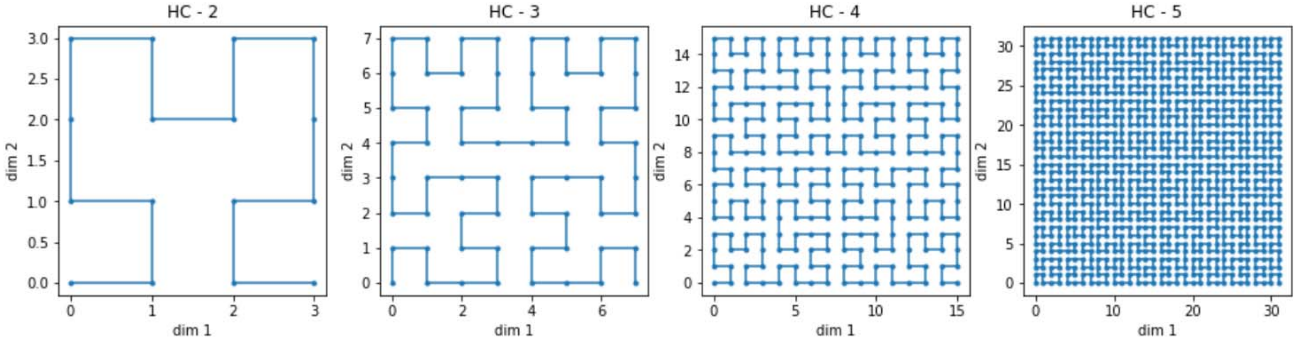


Fig. 2. Construction of approximations of the Hilbert curves of increasing τ=2; . . . ;5 in 2 dimensions.

to semantically cluster new uncategorised SDG data as well as new indicators with internationally yet unestablished methodology or standards and keep them close in the underlying physical networking environment of a distributed knowledge graph store. We introduce an algorithm that adopts HSFCs as the line of projection where new, gradually more refined, semantic categories are directly mapped onto. Our work proposes and experimentally corroborates the use of HSFCs to efficiently store distributed knowledge graph data, ensuring reduced access times and preservation of semantic closeness.

Section II describes the methodology we followed for use of an additional distributed environment layer based on HSFCs to map conceptually close, uncategorised according to existent SDG schema, data. First in the subsection II-A the proposed algorithm is theoretically founded and explained and an approach for data classification of entrant-indicators over this layer is described. Then, subsection II-B describes a detailed case study in a distributed knowledge graph environment, that experimentally evaluates our algorithm. Dataset and experimental setup are thoroughly discussed. The results are presented and discussed in light of theory along with the actual impact of the work in Section III. Section IV summarizes the paper's contribution and discusses future extensions of our work.

II. Methodology

A. Proposed algorithm

1) Hilbert Space-Filling Curves: A true Hilbert curve [5] is the limit of $\tau \to \infty$ of the τ^{th} discrete approximation to a Hilbert curve. HSFCs of 2 dimensions can be depicted on a NXN grid and the coordinates on the grid range in the space $x,\, y \in [0, N-1]$.

$$\text{N}=2^{\tau} \tag{1}$$

In Fig. 2 next order curve comprises of four gyrated reiterations of the previous order curve. In the next repetition, quadrants are split up into four sub-quadrants each and so on. The line is repetitively folded in such a way that passes by successive neighboring points without intersecting itself and with infinite iterations of the curve construction algorithm it will not omit any point on a continuous plane. HSFCs are always bounded by the unit square, with Euclidean length exponentially growing with τ. Continuity of the curve ensures that affinity of bins on the unit interval signifies affinity in the unit square as well. Two points $(x_1,\, y_1)$ and $(x_2,\, y_2)$ with affinity in HSFC of order τ_1 depict affinity in HSFC of order τ_2 >τ_1 as well. Hilbert approximations result in more efficient maintenance of local features as opposed to that achieved by linear ordering, while locality properties degrade with the increase of dimensions.

Previously, HSFCs [5] have been used along with Gray code and Z-order curves for heuristic multi-dimensional indexing via linearization. The wide spectrum of applications includes image compression, data visualization and peer-to-peer architectures [6]–[8]. HSFCs map multidimensional data to a single dimension maintaining spatial locality, namely affinity in the multidimensional space means relative affinity in the one-dimensional space. McSherry et al. [9] observed that edge ordering based on a HSFC substantially improves cache performance for single-threaded PageRank. Schmidt et al. [10] implemented a Distributed Hash Table (DHT)-based Web service discovery system leveraging HSFCs and mapped points of multidimensional space corresponding to service description components to DHT keys. Wang et al. [11] leveraged the spatial locality of HSFCs to store and display on request point-based spatial data in a spatial triple store.

2) Knowledge Graphs: A knowledge graph comprises of sets of triples that relate a subject entity to an object entity and encode domain and application knowledge. Knowledge graphs complimentarily serve for explainability that cognitively facilitates human-level intelligence. They serve for the representation of generic data interlinked by many relationships as well as for specific domains, such as biomedical research and manufacturing [12], [13]. They cover diverse application fields including search, data governance, question answering and recommendation. Distributed knowledge graphs integrate multiple and heterogeneous data sources, as their data are disseminated in a decentralised way across the web.

The impact of actions of organizations that harvest and process SDG data needs a specific context to be perceived in its wholeness. Their pursued goals and undertaken actions can be thematically interweaved and mutually influenced: one agent may pool interrelated existing content from federated repositories or one undertaken action may affect currently ongoing activities of other agents. Knowledge graphs are ideal for manifold interpretations of the societal implications of actions that apply to each SDG goal, for association of contributions of actions to concrete SDG targets and quantification of the exerted influence. Concerning SDGs, knowledge graphs: i) support explainable decisions and insightful recommendations, ii) measure the influence of users and communities and iii) improve the user experience, as they facilitate extraction and organization of knowledge in a distributed manner and serve for distributed knowledge matching.

Challenges associated with the uptake of distributed knowledge graph technologies include their efficient storage and use at scale [14]. Heretofore, SDG related systems have included in essence solely monitoring tools of SDGs data and metadata and mechanisms to enhance interoperability across independent information systems: UN [15] showcases use of mappings of terms to the UN Bibliographic Information System (UNIBIS) and the EuroVoc vocabularies.

We propose an algorithm (Alg. 1) for the efficient placement of Tier III SDG indicators in the underlying physical networking environment. The distinguishing element between the first two tiers is the fact that data of Tier I indicators are collected on a regular basis for not less than half of the countries and population in every relevant region [16].

Algorithm 1 Algorithm for insertion of SDG Tier III indicators in HSFCs

```
Input: $HSFC\_dims$, $HSFC\_order$, $indicator\_sentence$, $probe\_sentence$
Output: tuple $M$ = ($indicator$, $T$= ($x$ , $y$) $\in$ N)
Parameters: $bin$, $bin\_size$, $indicator$, $indicator\_number$, $T$=($x$,$y$) $\in$ N, Hilbert_Space_Filling_Curve $HSFC$
Initialisation :
$HSFC \leftarrow ConstructHSFC(\ HSFC\_dims,\ HSFC\_order)$
$bin\_size \leftarrow \frac{|q|}{|(2^{HSFC\_order})^2|-1}$
A $\leftarrow$compute embedding for $probe\_sentence$
for $indicator = 1$ to $indicator\_number$ do
    B $\leftarrow$compute embeddings for $indicator\_sentence[indicator]$
    compute $s$=$semantic\_similarity$ = $\frac{\cdot A \cdot B}{\|A\|\|B\|}$
    if $s(\ probe\_sentence,\ indicator\_sentence[indicator]) \leq Threshold_s$ then
        $bin \leftarrow \lfloor \frac{indicator}{bin\_size} \rceil$
        $T \leftarrow ObtainHSFCCoordinates(\ bin,\ HSFC)$
    end if
end for
```

Differently from Tier I and II, Tier III indicators are not associated with any existent methodology / standards. The order of the HSFCs used defines the range of possible coordinates. We incorporate a binning mechanism to ensure that each new indicator can be projected to a tuple of coordinates in the higher dimension space. Bins hold consecutive elements of the data vector. The suggested mapping represents the indexing mechanism for the data in the distributed knowledge graph storage prototype we develop.

The first layer of the distributed knowledge graph store will entail semantic representation of data. In the next layer, which acts as a substrate of the network topology, we split up the indexing area in semantically homogenous areas through HSFCs. Use of curves in this building block proves beneficial for preserving the neighbourhood property of concepts expressed by the indicators, as semantically related terms, more probable to respond to a user query, will be placed in the vicinity. In our suggestion linearization is implemented as an overlay upon existing unidimensional search structures and the distributed file system, that ensures distribution and sharding that scale. Multidimensional queries upon the distributed knowledge graph can be mapped to unidimensional queries, that range from the minimum to maximum linearization points of the initial query (Fig. 3).

We are interested in relative positioning that expresses affinity. The reverse process of HSFCs mapping, when a position in the description space for higher-order partitioning needs to be translated into a position in the indicators vector, is not applicable and does not cause any issue in our scenario. The algorithm can be further modified to scale with new entries in terms of targets, goals and other potential refinements of the SDG ontology with corresponding increase in the order of the HSFC.

3) Similarity Assessment: We quantify the semantic textual similarity of each probe sentence, that is a candidate entrant-indicator, with existent SDG indicators ($indicator$). In this direction, we compute semantic textual similarity as calculated in Sentence-BERT (SBERT) [17], an approach that extracts and compares semantically meaningful sentence embeddings.

- Each word in the $sentence_i$ is preprocessed (e.g., tokenized).
- Each processed word in the sentence is encoded into vectors v_{ij} of 300 dimensions. Word2vec encodes similar words closer to each other in the vector space.
- To derive vector representation for $sentence_i$: average of such v_{ij} vector representations for $j = \{i, w\}$ is computed, where w is the number of words in the sentence.
- Sentence embeddings for all existent indicators in the SDG taxonomy are precalculated. They are assumed to be close in the 300 dimensional vector space if they are similar.
- Thus, computing cosine similary between the (300 dim) vector representation provides ideal score of 1, if the sentences are identical and score of 0, if the sentences are maximally dissimilar to each other.

4) Data Classification of Entrant-Indicators: We aim to avoid unnecessary congestion of specific subquadrants in the HSFC mapping, thus we team up semantically close entrants, namely indicators of Tier III. In this direction, we aim to

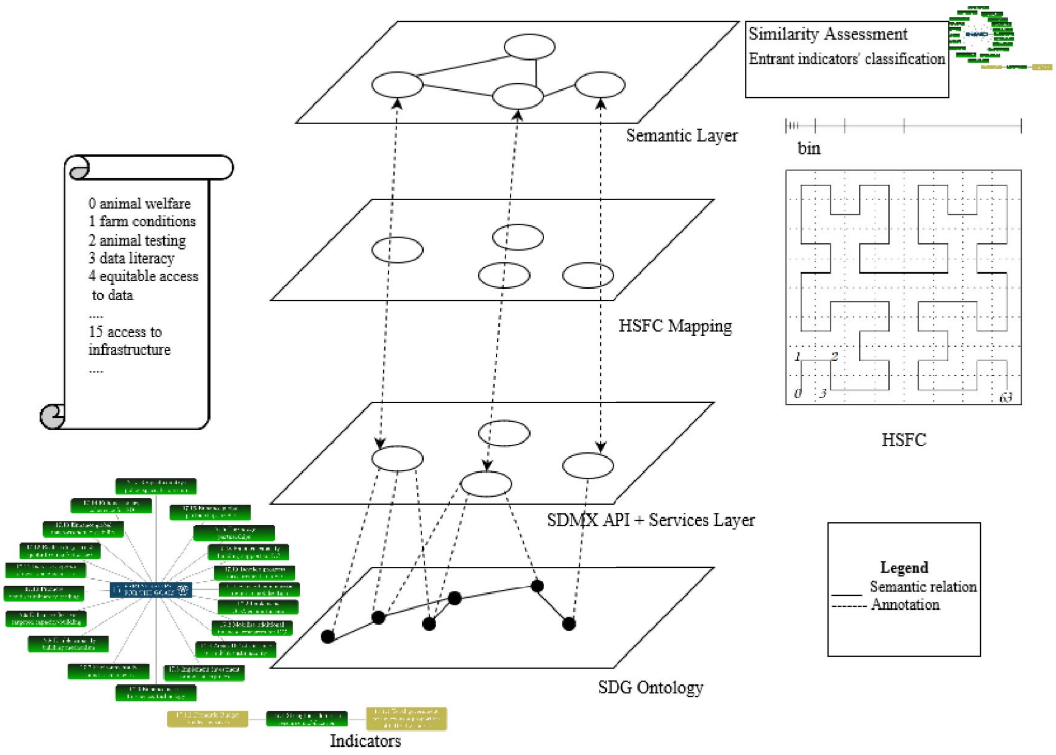


Fig. 3. Suggested implementation framework of the algorithm.

categorize points in $\mathbb{R}^n$ without exploiting SDG schema features. For the data classification, we use a dataset of N_{DC} individuals described by Q categorical variables and construct the $N_{DC} \times J$ indicator matrix Z, where:

$$J = \sum_{q=1}^{Q} J_q \tag{2}$$

rows denote the datasources, namely nodes of the graph store where data associated with the indicators reside, and columns denote the indicators of the uncategorised SDG data. We calculate a matrix of proportions P where $p_{ij} = n_{ij}/n$ and n is the sample size, summing up all values of N_{DC}. r and c are the sums along the rows and along columns respectively.

Categorization is based on chi-squared distances between two entrant-indicators:

$$\text{dist}^2_{\chi^2}(ind_j, ind_{j'}) = \sum_{i=1}^{N_{DC}} \frac{1}{r_i}\left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}}\right)^2 \tag{3}$$

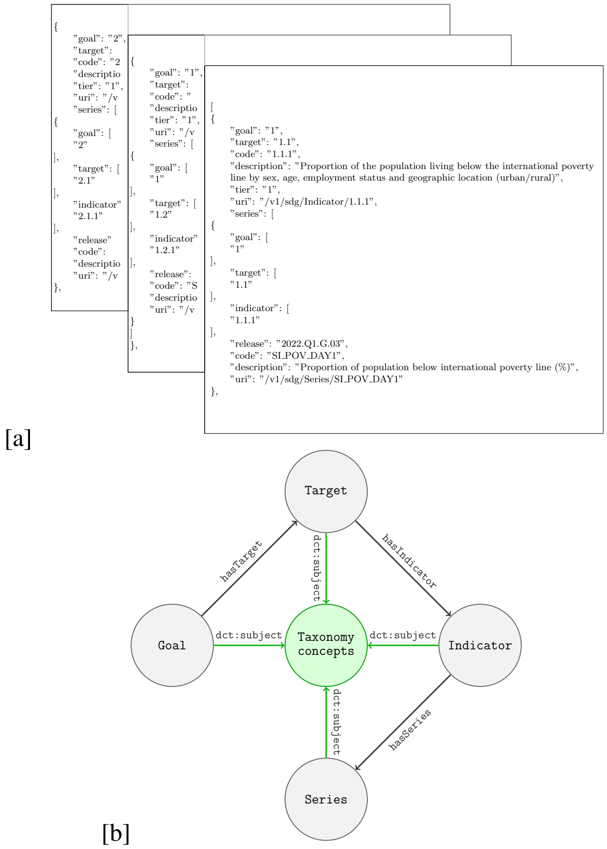
The distance is reduced when there is overlapping between individuals belonging to multiple categories. Our aim is to project the points onto a subspace of lower dimensionality, within which the eigenvectors u_k are the result of eigenvalue decomposition of $PD_c^{-1}P^TD_r^{-1}$. So we solve the equation:

$$\frac{1}{Q}ZD^{-1}Z^Tu_k = \lambda_k u_k \tag{4}$$

where Z is the indicator matrix, D_r, D_c the diagonal matrix of row and column masses respectively.

B. Case study

1) Dataset: We harvested a dataset of 2,21M. entries in total that includes all dimensions (standard demographic info, the whole variety of different age profiles, etc.), time periods and area codes, described through the UNM49 standard available for each indicator from 2000 onwards. We used the API of UN Statistics Division [18] with a set of scripts written in TypeScript and ran in the node.js environment. Dataset was particularly focused on indicators and list of all available SDG indicators was our starting point in the API, providing all available indicators in a self-contained response. Within the indicator related datasets, we collected 3 core datasets, while others were mostly redundant data provided for different data access or interpretation. Our dataset IndicatorData includes 169 targets, 248 indicators (with 13 replicated under two/three different targets), as they were described in the 2022 refinement of the SDGs, as well as 663 data series for the quantification of the SDGs' accomplishment [3]. Since 2022 the classification entails 136 indicators of Tier I, 91 indicators of Tier II and 4 indicators consisting of modules of disparate tiers [16]. The dataset includes series information and goal - target hierarchy with overall 663 series across 248 indicators (Fig. 4). The number of data entries per each indicator is 4150 after removal of 20% of top and tail outliers. There are multiple properties describing each observation (data entry) and their



```
[
{
    "goal": "1",
    "target": "1.1",
    "code": "1.1.1",
    "description": "Proportion of the population living below the international poverty
    line by sex, age, employment status and geographic location (urban/rural)",
    "tier": "1",
    "uri": "/v1/sdg/Indicator/1.1.1",
    "series": [
{
    "goal": [
    "1"
],
    "target": [
    "1.1"
],
    "indicator": [
    "1.1.1"
],
    "release": "2022.Q1.G.03",
    "code": "SI_POV_DAY1",
    "description": "Proportion of population below international poverty line (%)",
    "uri": "/v1/sdg/Series/SI_POV_DAY1"
},
```

Fig. 4. (a) IndicatorData dataset indicative excerpts. (b) SDG schema of the dataset.

Statistical Data and Metadata eXchange (SDMX)-standardized code equivalents are also provided. Table I depicts number of occurences of uncategorised indicators in graph store nodes and Table II describes uncategorised entrant indicators, that are not included in the dataset we harvested and do not follow SDG schema beforehand.

PivotData dataset returns a list of observations pivoted by year. This dataset contains all data described in IndicatorData aggregated for the whole observation period and showing only pivoting years in the years property of each data entry. The property was serialized and we deserialized it for the convenience of data manipulation.

There are 247.251 entries in total, with 550 entries per indicator after removal of 20% of outliers (top and tail).

2) Experimental Setup: We evaluate our algorithm in an experimental distributed environment over a key-value store of SDG data, that we collected. We use multiple servers and Hypertext Preprocessor (PHP) clients as APIs to handle cached values in a scheme built on Memcached, an optimized distributed hash map-based mechanism (Fig. 5). Placement of data with HSFCs is compared to default placement scheme of the prototype distributed cache mechanism in terms of response time for the executed SELECT queries and in terms of disk I/O. Experimental setup settings are described in Table III. In order to make clusters for entrant indicators and put their content in close Hilbert areas, we use the Agglomerative Hierarchical Clustering (AHC).

TABLE I
NUMBER OF OCCURRENCES OF UNCATEGORISED TIER III ENTRANT INDICATORS IN DATASOURCES (GRAPH STORE NODES).

Table	**Indicators**											
Datasource	***I1***	***I2***	***I3***	***I4***	***I5***	***I6***	***I7***	***I8***	***I9***	***I10***	***I11***	***I12***
1	16	16	16	16	17	17	16	16	16	16	16	16
2	15	18	18	14	16	16	15	16	17	16	17	15
3	16	17	17	15	16	16	14	17	18	16	17	16
4	14	18	16	14	16	18	16	16	17	17	18	16
5	16	16	16	16	16	17	16	16	15	16	16	16
6	15	17	15	15	15	17	15	16	17	16	18	15
7	16	17	17	15	15	16	15	18	17	17	17	15
8	14	18	17	14	15	16	16	16	18	16	18	15
9	16	16	15	16	17	18	16	15	16	16	16	16
10	16	16	15	16	16	17	16	16	15	17	16	16

TABLE II
TIER III ENTRANT INDICATORS

Indicator	***Description***
I1	inclusive access to knowledge
I2	abolish unnecessary animal testing
I3	stop animal caging
I4	access to infrastructure
I5	cross border processing
I6	data erasure
I7	data portability
I8	data literacy
I9	improve animal welfare
I10	promote research
I11	improve farm conditions
I12	equitable access to knowledge

The similarity threshold we choose is minimum to allow augmentation of data with the whole set of entrant indicators. AHC proceeds with combination of clusters from the simple level of clusters-individuals to merging pairs of them with a bottom-up approach. The metric used in our setup is the Euclidean distance for pairwise observations.

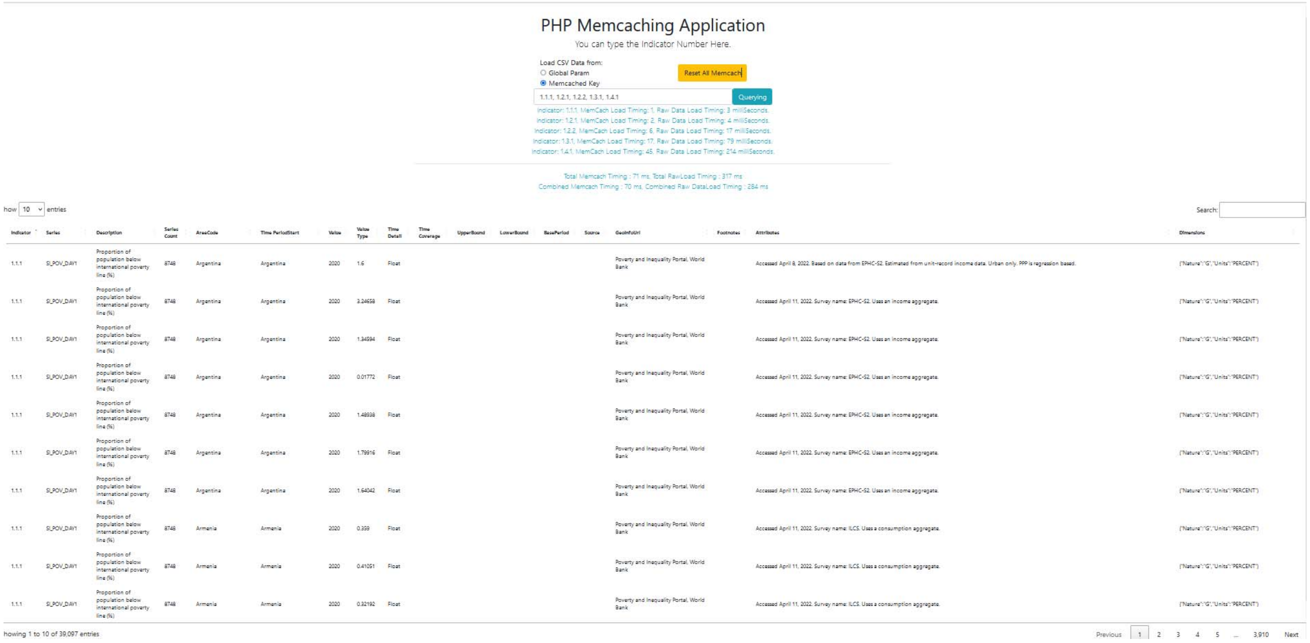
III. RESULTS

A. Cost-aware Data Classification of Entrant Indicators

Firstly, the approach for cost-aware data classification of entrant-indicators is verified. Concerning the uncategorised indicators of Table II, cutting the dendrogram (Fig. 6) at the

TABLE III
EXPERIMENTAL SETUP

Parameter	***Value***
Dataset	2,21M. entries
Number of servers	3
Queries	SELECT
HSFC dimensions	2
HSFC order	3
Memcached server chunk size	1MB
Memcached server page size	40



PHP Memcaching Application

You can type the Indicator Number Here.

Load CSV Data from:
○ Global Param
● Memcached Key

Reset All Memcach

1.1.1, 1.2.1, 1.2.2, 1.3.1, 1.4.1

Querying

Indicator: 1.1.1, MemCach Load Timing: 1, Raw Data Load Timing: 3 milliSeconds.
Indicator: 1.2.1, MemCach Load Timing: 2, Raw Data Load Timing: 4 milliSeconds.
Indicator: 1.2.2, MemCach Load Timing: 6, Raw Data Load Timing: 17 milliSeconds.
Indicator: 1.3.1, MemCach Load Timing: 17, Raw Data Load Timing: 79 milliSeconds.
Indicator: 1.4.1, MemCach Load Timing: 45, Raw Data Load Timing: 214 milliSeconds.

Total Memcach Timing : 71 ms, Total RawLoad Timing : 317 ms
Combined Memcach Timing : 70 ms, Combined Raw DataLoad Timing : 284 ms

how 10 entries

Search:

Indicator	Series	Description	Series Count	AreaCode	Time PeriodStart	Value	Value Type	Time Detail	Time Coverage	UpperBound	LowerBound	BasePeriod	Source	GeoInfoUrl	Footnotes	Attributes	Dimensions
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Argentina	Argentina	2020	1.6	Float						Poverty and Inequality Portal, World Bank		Accessed April 8, 2022. Based on data from EPHC-S2. Estimated from unit-record income data. Urban only. PPP is regression based.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Argentina	Argentina	2020	3.24658	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: EPHC-S2. Uses an income aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Argentina	Argentina	2020	1.34594	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: EPHC-S2. Uses an income aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Argentina	Argentina	2020	0.01772	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: EPHC-S2. Uses an income aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Argentina	Argentina	2020	1.48938	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: EPHC-S2. Uses an income aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Argentina	Argentina	2020	1.79916	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: EPHC-S2. Uses an income aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Argentina	Argentina	2020	1.64042	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: EPHC-S2. Uses an income aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Armenia	Armenia	2020	0.359	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: ILCS. Uses a consumption aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Armenia	Armenia	2020	0.41051	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: ILCS. Uses a consumption aggregate.	{"Nature":"G","Units":"PERCENT"}
1.1.1	SI_POV_DAY1	Proportion of population below international poverty line (%)	8748	Armenia	Armenia	2020	0.32192	Float						Poverty and Inequality Portal, World Bank		Accessed April 11, 2022. Survey name: ILCS. Uses a consumption aggregate.	{"Nature":"G","Units":"PERCENT"}

howing 1 to 10 of 39,097 entries

Previous 1 2 3 4 5 … 3,910 Next

Fig. 5. An experimental distributed environment over a key-value store of SDG data based on the hash map-based mechanism of Memcached.

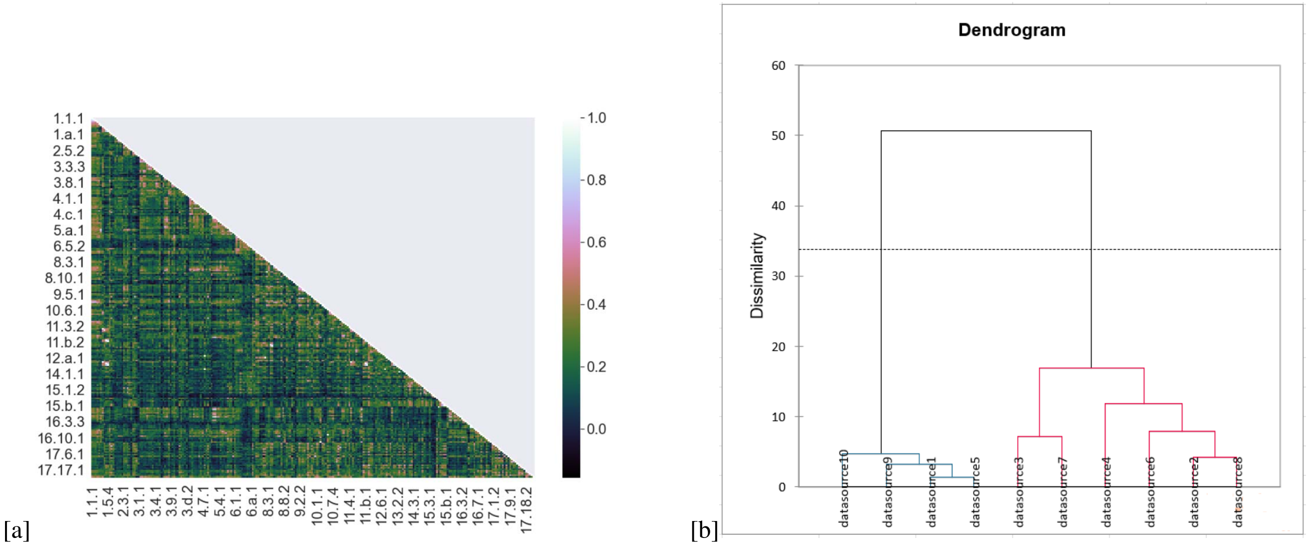


Fig. 6. (a) Similarity among existent SDG indicators. (b) Entrant indicators' AHC dendrogram based on their number of occurences in datasources.

height of the dotted line verifies a coarser clustering of two semantic categories, namely of datasources (1,5,9,10) covering topics (1,4,5,6,7,8,10,12) associated with data and those of the rest datasources (2,3,4,6,7,8) associated with animal issues (2,3,9,11). Explicit reference to terms (”animal”, ”data”) here is irrelevant. Thus, datasources (1,5,9,10) and datasources (2,3,4,6,7,8) should be put in two separate subquadrants in the Hilbert unit square. As for existent SDG indicators, Fig. 6 depicts their comparison in terms of semantic similarity.

The Principal Component Analysis (PCA) scree plot indicates that two dimensions F1, F2 suffice for the visual interpretation of the analysis, since the sum of first two eigenvalues is sufficient percentage of variance. The quality of the fit is measured by the percentage of inertia related to the two-dimensional map, namely the ratio of variance of coordinates of individuals on the axis to the total variance of coordinates of individuals. The quality is high for our dataset of restricted size (10 individuals (datasources) and 14 categorical variables (indicators)) and high data interlinking. With the eigenvalue λ_d equal to the variance of the points of each indicator on d-dimension:

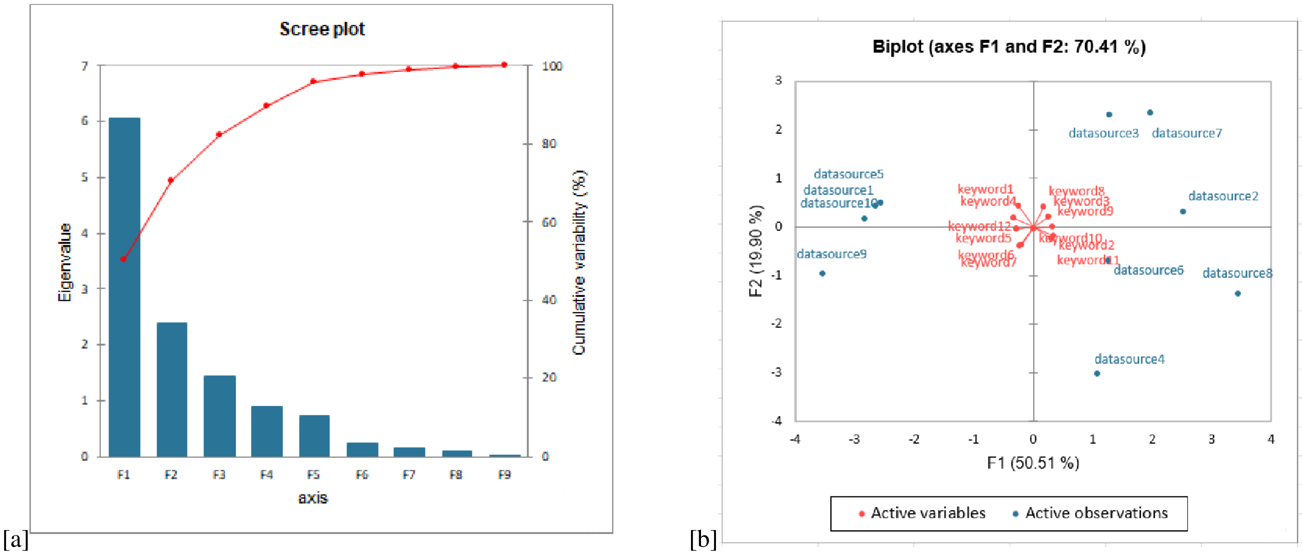


Fig. 7. PCA of new indicators. (a) Scree plot with first two axes F1, F2 contributing. (b) Biplot PCA denoting the suggested division of layer to two Hilbert areas.

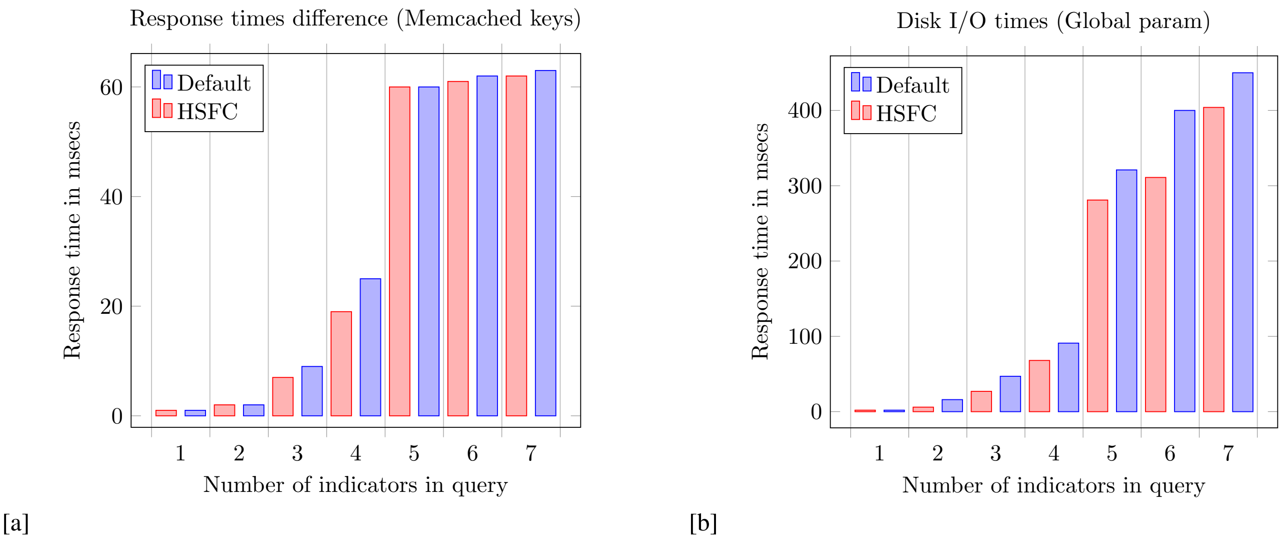


Fig. 8. (a) Response time differences for HSFC mapping assigned to Memcached keys and (b) Disk I/O times for HSFC mapping passed as Globalparam.

$$(\lambda_1 + \lambda_2) \ / \sum_{d=1}^{9} \lambda_d = 70,41\% \qquad (5)$$

the biplot verifies the split up of the datasources to two main semantic categories with active observations corresponding to the selected datasources and active variables corresponding to selected indicators (Fig. 7).

B. Response Time Reduction

We ran multiple sets of queries in an experimental distributed environment over a key-value store of SDG data with multiple servers and PHP clients as APIs to handle cached values in a scheme built on Memcached. After each set of queries the Memcached server was reset. We notice significant reduction in average response times for selection queries of combined indicators. Time difference between HSFC storage scheme and baseline distributed key-value store approach is more obvious in the case of disk I/O times (Global parameters used). There is also improvement in response times when HSFC mapping is loaded into Memcached keys directly, which is more obvious for combinations of sets of up to 4 indicators in our setup (Fig. 8). The improvement in terms of memory response times can be further increased with further paging configuration, due to the nature of Memcached custom memory manager (slabs hold objects within specific ranges and slabs contain pages, split up in chunks) and the fact that a single indicator's entries reach up to 20MBs in our detailed dataset.

In light of HSFC theory, the locality in the multidimensional space describing the semantically associated indicators indicators is preserved after their mapping, as input items with higher semantic similarity are mapped to nearby addresses. Hence, nearby mapping is leveraged and the placement of conceptually close SDG indicator data on an HSFC as the line of projection indeed reduces retrieval times. The suggested topological mapping scheme is nondisruptive in terms of space and maintains local feature correlations of the original space. HSFC points were coarsely equivalent to servers in the experimental distributed environment. Therefore, further refinement at a graph store node level and per server could lead to even better results, because communication cost among servers would be alleviated.

The practical impact of our work is that data retrieval times are reduced for semantically close data, that have not been categorised according to the prevailing schema. Our approach empowers SDG knowledge graphs for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions, as data are accessed in reduced retrieval times. It facilitates quicker measurement of influence of users and communities on specific goals and serves for faster distributed knowledge matching, as semantic cohesion is preserved.

IV. Conclusions and Discussion

Our work aims to support the efficient processing of SDG data and the seamless integration of novel indicators. An efficient storage scheme is needed for new uncategorised SDG data as well as indicators with internationally yet unestablished methodology and standards. In this paper, we introduce a mapping method based on HSFCs as the line of projection where semantic categories of conceptually close SDG indicator data, uncategorised according to the existent schema, are directly mapped onto. A case study on real SDG data in a distributed knowledge graph store validates that data retrieval time is reduced. The proposed algorithm can be adapted for targets, goals, and potential future refinements of the SDG ontology.

Our approach empowers SDG knowledge graphs for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions, as data are accessed in reduced retrieval times. It facilitates quicker measurement of influence of users and communities on specific goals and serves for faster distributed knowledge matching, as semantic cohesion of data is preserved.

Future extension includes scaling our approach with the introduction of load balancing mechanisms at runtime or periodical batch-level processing of data, which will ensure that in case of skewed distributions (more occurences of specific indicators or semantic categories) the equivalent subquadrants in the HSFC unit square will not be congested. We aim, furthermore, to study how increasing order of HSFCs affects performance. In another direction, we intend to explore geolocation features of indicators to leverage multiple HSFCs for spatial joins and range queries, as well as optimize queries to correspond to global search trends on SDG data.

The collective effort to optimally harmonize sustainability goals requires the conscious technological contribution of sustainable IT for timely and reliable data. Our work aspires to contribute in this direction and prove useful for practitioners gathering and assessing SDG data, including intergovernmental organizations, government agencies and social welfare organizations.

References

[1] UN, "A/res/70/01," October 2015. [Online]. Available: https://undocs.org/

[2] ——, "Sustainable development goals," September 2015. [Online]. Available: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

[3] ——, "Global SDG indicator framework after 2022 refinement," 2022. [Online]. Available: https://unstats.un.org/sdgs/indicators/indicators-list/

[4] ——, "SDG taxonomy," November 2019. [Online]. Available: http://metadata.un.org/sdg/

[5] D. Hilbert, "Über die stetige abbildung einer linie auf ein flächenstück," in *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes*. Springer, 1935, pp. 1–2.

[6] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the hilbert space-filling curve," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 1, pp. 124–141, 2001.

[7] J. K. Lawder and P. J. King, "Using space-filling curves for multi-dimensional indexing," in *Proc. British National Conference on Databases '00*. Springer, 2000, pp. 20–35.

[8] M. Ammari, D. Chiadmi, and L. Benhlima, "A semantic layer for a peer-to-peer based on a distributed hash table," in *Proc. Int. Conf. on Informatics Engineering and Information Science (ICIEIS) '11*. Springer, 2011, pp. 102–114.

[9] F. McSherry, M. Isard, and D. G. Murray, "Scalability! but at what COST?" in *Proc. 15th USENIX Conf. on Hot Topics in Operating Systems (HotOS XV)*, May 18-20, 2015. [Online]. Available: https://www.usenix.org/system/files/conference/hotos15/hotos15-paper-mcsherry.pdf

[10] C. Schmidt and M. Parashar, "A peer-to-peer approach to web service discovery," in *Proc. World Wide Web (WWW) '04*, vol. 7, no. 2. Springer, 2004, pp. 211–229.

[11] C.-J. Wang, "Database indexing for skyline computation, hierarchical relational database, and spatially-aware sparql evaluation engine," Ph.D. dissertation, 2015.

[12] A. Santos, A. R. Colaço, A. B. Nielsen, L. Niu, M. Strauss, P. E. Geyer, F. Coscia, N. J. W. Albrechtsen, F. Mundt, L. J. Jensen, and M. Mann, "A knowledge graph to interpret clinical proteomics data," *Nature Biotechnology*, vol. 40, p. 692–702, 2022.

[13] K. S. Aggour, V. S. Kumar, P. Cuddihy, J. W. Williams, V. Gupta, L. Dial, T. Hanlon, J. Gambone, and J. Vinciquerra, "Federated multimodal big data storage & analytics platform for additive manufacturing," in *Proc. IEEE Big Data '19*, Los Angeles, CA, USA, Dec. 9-12, 2019, pp. 1729–1738.

[14] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022.

[15] UN, "Linked sdg," 2022. [Online]. Available: https://linkedsdg.officialstatistics.org/

[16] ——, "Tier classification for global SDG indicators," June 2022. [Online]. Available: https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/

[17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. on Emp. Methods in Nat. Lang. Processing and the 9th Int. Joint Conf. on Nat. Lang. Processing (EMNLP-IJCNLP) '19*. Hong Kong, China: ACL, November 2019, pp. 3982–3992.

[18] UN, "SDG API." [Online]. Available: https://unstats.un.org/SDGAPI/swagger/