LISE: a Logic-based Interactive Similarity Explainer

Simona Colucci[1], Francesco Maria Donini[2] and Verdiana Schena[1]

[1]*Politecnico di Bari, Via Orabona 4, 70125, Bari, Italy*
[2]*Università della Tuscia, Via S. Maria in Gradi 4, 01100 Viterbo, Italy.*

Abstract

This work presents LISE (Logic-based Interactive Similarity Explainer), a system for explaining the similarity of clusters of RDF resources, by identifying common characteristics in their RDF descriptions. LISE follows a pipeline that consists of four main modules: *Machine Learning Module*, which creates a representation of RDF resources as vector embeddings and clusters them; *Logic-Based Module*, which, for each cluster, computes a Knowledge Graph (with blank nodes) modeling the common characteristics of resources in the cluster; *Natural Language Generation Module*, which translates the computed Knowledge Graphs into human-readable descriptions; and *User Interaction and Feedback Loop*, which collects user feedback about the relevance of generated explanations. LISE operates in a closed loop, leveraging user feedback to refine embeddings and subsequently improve clustering. It was tested on an RDF dataset containing structured drug-related information, demonstrating promising results in terms of explainability and interpretability of clustering results.

Keywords

Knowledge Graph Embeddings, Resource Description Framework, Interactive Clustering, User Interaction, Natural Language Generation, Explainable Artificial Intelligence

1. Introduction

Unsupervised learning, particularly clustering, is a fundamental technique for the initial exploration of unstructured data. Clusterization groups similar data points together, enabling users to identify hidden patterns and potential classifications by analyzing these groups. However, the notion of "similarity" is inherently subjective and depends on the users' objectives and the specific features they prioritize[1]. For instance, in drug-related data, different aspects such as protein interactions or sequence similarity might be of primary importance[2]. Once a clustering algorithm generates groups, it is essential to understand the characteristics that define each cluster. This is particularly relevant in interactive clustering [1], where users iteratively refine the process based on their needs. However, comprehensible feature extraction is not always straightforward, especially when dealing with purely numerical data or when items are identified by IRIs, whose features can be inferred from a linked Knowledge Graph (KG). While KGs can theoretically provide meaningful feature information, the embedding process used in clustering often obscures the shared characteristics within a cluster. To address this challenge, we developed LISE (Logic-based Interactive Similarity Explainer), a modular system that clusters groups of RDF resources, and for each cluster, computes an RDF KG describing the commonalities of resources and generates an English description from such KG; finally, system users vote on the relevance of every single part of the description to improve the clustering process in an interactive feedback loop.

The natural language description of cluster commonalities serves as an explanation service for the clustering method. From the perspective of eXplainable Artificial Intelligence (XAI) [3], LISE can be classified as *text-based*, *method-agnostic* (compatible with any clustering approach), and *post-hoc* (providing explanations after clustering is completed).

With respect to previous work, LISE introduces two main innovations: the application of pruning techniques to filter out irrelevant information, detailed in Section 2.3, and the inclusion of user interactivity, which enables a feedback loop to iteratively refine the clustering process.

SeWebMeDa-2025: 8th International Workshop on Semantic Web solutions for large-scale biomedical data analytics, June 1, 2025, Portorož, Slovenia

simona.colucci@poliba.it (S. Colucci); donini@unitus.it (F. M. Donini); v.schena@phd.poliba.it (V. Schena)

The paper is organized as follows. In the next section, LISE architecture and main modules are described, with reference to an extended use case addressing drugs comparison problem. In particular, Section 2.1 explains how the input KG is extracted and how entities are selected from the reference RDF dataset: DrugBank[1]. Sections 2.2–2.5 explain the working mode of each system component. Section 3 closes the paper.

2. System

LISE explains the similarity of RDF resources grouped through clustering by logically computing commonalities in their RDF descriptions. It implements a pipeline that transforms an RDF dataset into a KG, generates a representation for KG entities based on vector embeddings, clusters embeddings, computes a logic-based explanation for clusters, and iterates the process based on user feedback.

LISE architecture is depicted in Figure 1 and structured into four main modules:

- *Machine Learning Module*: Responsible for representing RDF resources as feature vectors by training RDF2Vec [4] embedding models and performing clustering to group resources based on similarity.
- *Logic-Based Module*: Computes an RDF KG that models the common characteristics of the resulting clusters while filtering out irrelevant information.
- *Natural Language Generation Module*: Converts the RDF KG into a human-readable description by using a template-based approach.
- *User Interaction and Feedback Loop*: Collects user feedback on each explanation, leveraging this feedback to refine and retrain the embeddings, thereby improving alignment with user-defined relevance criteria.

Each module plays a crucial role in enhancing the quality of explanations, contributing to a feedback loop that iteratively improves the interpretability of clustering results.

The following sections provide a discussion of system functionalities, by introducing the role of each module. Preliminarily (Section 2.1), the entity selection process is introduced. Then, Section 2.2 details the generation of embeddings from RDF resources (Section 2.2.1) and the application of clustering (Section 2.2.2). Section 2.3 describes the logic-based computation process that produces, for each cluster, an RDF KG modeling characteristics shared by cluster items. Section 2.4 explains how LISE generates natural language descriptions from the Knowledge Graphs obtained in Section 2.3; additionally, it explores the integration in LISE of a Large Language Model (LLM) for Natural Language Generation (NLG). Finally, Section 2.5 illustrates how user feedback is collected and used to iteratively improve the system.

2.1. Knowledge Graph and Entity Selection

We show LISE capabilities on DrugBank, a dataset containing structured information about drugs and available in RDF format. The dataset is first converted into a pyRDF2Vec[2] KG[3], in which the set of drugs to divide into clusters is selected by choosing entities belonging to the Drug class and characterized as small molecules. This selection process results in a dataset of 7,391 resources. To enhance the quality of the graph and eliminate redundant information, all predicates irrelevant to drug comparison are excluded. In our examples, we identified a preliminary list of so-called "stop-patterns"[5], that we list in Appendix A, for the sake of reproducibility. For each drug, a subgraph is extracted from the initial KG, up to a maximum depth of 7. These subgraphs are employed by the Logic-based Module, which is further analyzed in Section 2.3.

[1] https://download.bio2rdf.org/files/current/drugbank/drugbank.html

[2] https://pyrdf2vec.readthedocs.io/en/latest/index.html

[3] pyRDF2Vec KG is a structure for modeling and managing RDF data in the form of a Knowledge Graph, in which triples are represented as vertices and edges, enabling the extraction of RDF paths.

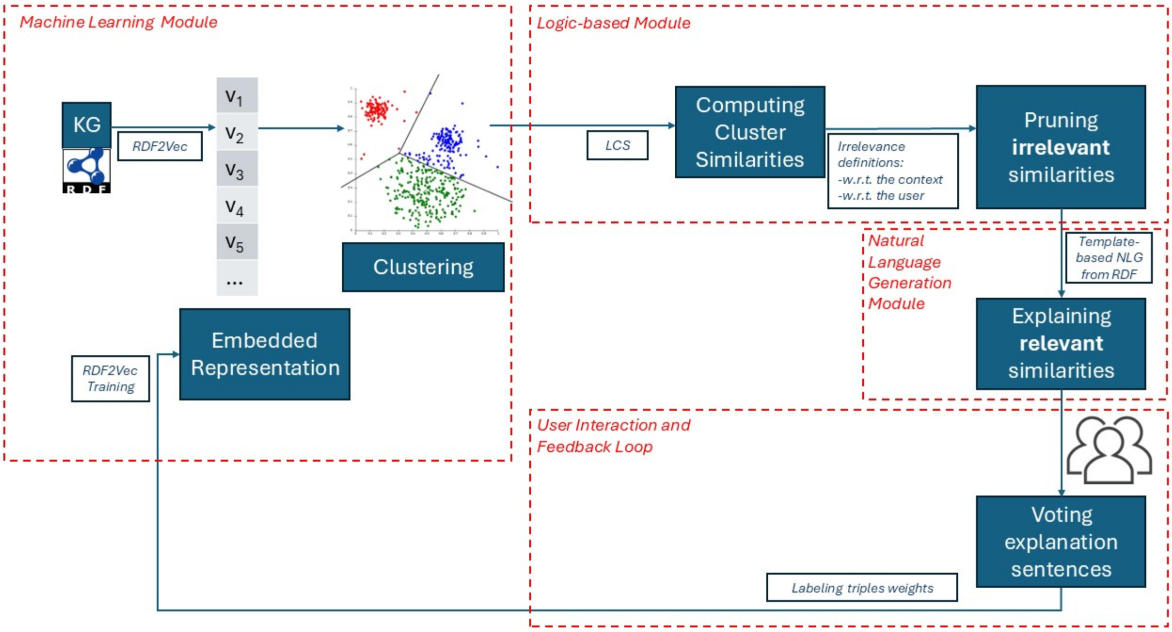

Machine Learning Module
KG
RDF
RDF2Vec
v_1
v_2
v_3
v_4
v_5
...
Clustering
RDF2Vec Training
Embedded Representation
Logic-based Module
LCS
Computing Cluster Similarities
Irrelevance definitions:
-w.r.t. the context
-w.r.t. the user
Pruning irrelevant similarities
Natural Language Generation Module
Template-based NLG from RDF
Explaining relevant similarities
User Interaction and Feedback Loop
Voting explanation sentences
Labeling triples weights

Figure 1: Architecture of LISE. Each module is enclosed by a dashed red border. Blue rectangles describe activities, and captions near arrows describe methods at the basis of the switch from one activity to the next one.

2.2. Machine Learning Module

2.2.1. Embedding generation

To represent entities as numerical feature vectors, we employ RDF2Vec [4], a method that applies the word embedding principle to KGs. RDF2Vec represents each entity as a point in a continuous vector space, providing versatile input for various machine learning applications.

Specifically, we use pyRDF2Vec, a Python implementation that extracts walks from KGs to generate embeddings. Since the number of possible walks can be exponential in the worst case, pyRDF2Vec samples possible walks, allowing for the selection of different sampling strategies [6]. RDF2Vec operates by generating sequences of entities and relations (walks) within the KG, treating them as sentences from a text corpus. These sequences are then used to train a Word2Vec[4] model, which learns vector representations for each entity, capturing both semantic and structural relationships. The embedding generation process follows these key steps:

- *Definition of the walking strategy*: A random walking strategy is employed, where each entity in the KG is explored through paths with a maximum depth of 7. The number of walks generated per entity is limited to 3310, corresponding to the highest number of triples present in the extracted subgraphs. The random state parameter is set to 42.
- *Definition of the sampling strategy*: We build upon the existing *Predicate Frequency Weight* [7] strategy, which assigns a weight to each predicate based on its frequency of occurrence in the KG, but we address a more specific concept of weight — according to the call in the pyRDF2Vec documentation to develop new walking, sampling, and embedding strategies. In particular, we introduce a custom sampling strategy, namely *Predicate Relevance Weight*, which assigns weights to specific walks based on their relevance for cluster explanation. These scores are derived from user feedback, where users rate the relevance of explanatory sentences generated for each cluster (see Figure 9 for a screenshot). LISE learns these weights through an interactive component, employing a regression model to predict the weights. In the first iteration, the

[4] https://radimrehurek.com/gensim/models/word2vec.html

weights for each predicate are manually set to 0.01 or 0.02, with the exception of the predicate `http://bio2rdf.org/drugbank_vocabulary:category`, which is assigned a weight of 0.99. This is because our goal was to cluster resources based on this specific predicate, and therefore, we gave it a higher weight.

- *Definition of the embedding strategy*: The extracted walks are processed as textual sentences and used as input for the Word2Vec model. The model is configured with the skip-gram architecture while maintaining default hyperparameters for training.

After applying RDF2Vec to the entire KG, each entity is transformed into a fixed-dimensional vector. These vector representations enable the application of clustering techniques to group semantically related entities.

While RDF2Vec provides compact and interpretable embeddings, their effectiveness depends on how well they preserve entity semantics within the given domain context. Although RDF2Vec generates embeddings that are computationally efficient in terms of size, their ability to retain the original semantic content of RDF entities remains disputable [8]. Specifically, previous studies have demonstrated that the assumption underlying RDF2Vec—that similar entities will have similar embeddings—is not consistently supported in real-world machine learning applications [9]. Indeed, when analyzing the common properties of the entities within the clusters using LISE, we find that the explanations are often uninformative. This indicates that, although clustering groups entities that are close in the embedding space, they are not necessarily semantically similar. To analyze the distribution of embeddings, we employ Principal Component Analysis (PCA) from the scikit-learn library [10] to project the original 100-dimensional embeddings into a two-dimensional space for visualization (Figure 2).

2.2.2. Clustering

Clustering is performed using the k-means algorithm via the scikit-learn library [10] with the parameters in Table 1. Clustering models alternative to k-means were tested, but they did not produce significant improvements in terms of amount or relevance of common information within each cluster.

Number of Clusters	302
Initialization	k-means++
Random State	42
Number of Initializations	5

Table 1
K-means clustering parameters.

Figure 3 shows the complete clustering results visualized in 2D with PCA. To evaluate the quality of the clustering in terms of both inter-cluster separation and intra-cluster cohesion, the average Silhouette Score is computed considering all 302 clusters. This results in a value of 0.1415, which is close to 0, indicating that there is an overlap between the clusters. Since the visualization does not exhibit a clear spatial separation between clusters, the 10 most cohesive clusters – identified by evaluating the average Euclidean distance between each cluster's points and its centroid – each containing at least two elements, along with their centroids, are shown in Figure 4.

2.3. Logic-Based Module

LISE explains clusters by performing a logic-based processing of RDF KGs, for computing the characteristics shared among clustered resources. A core component of this process is the computation of the *Least Common Subsumer* (LCS) [11], an RDF KG which describes the common features of a group of RDF resources. The LCS is computed on the subgraphs previously extracted for the resources to compare (Section 2.1).

To ensure meaningful graph comparison, we perform some optimizations:

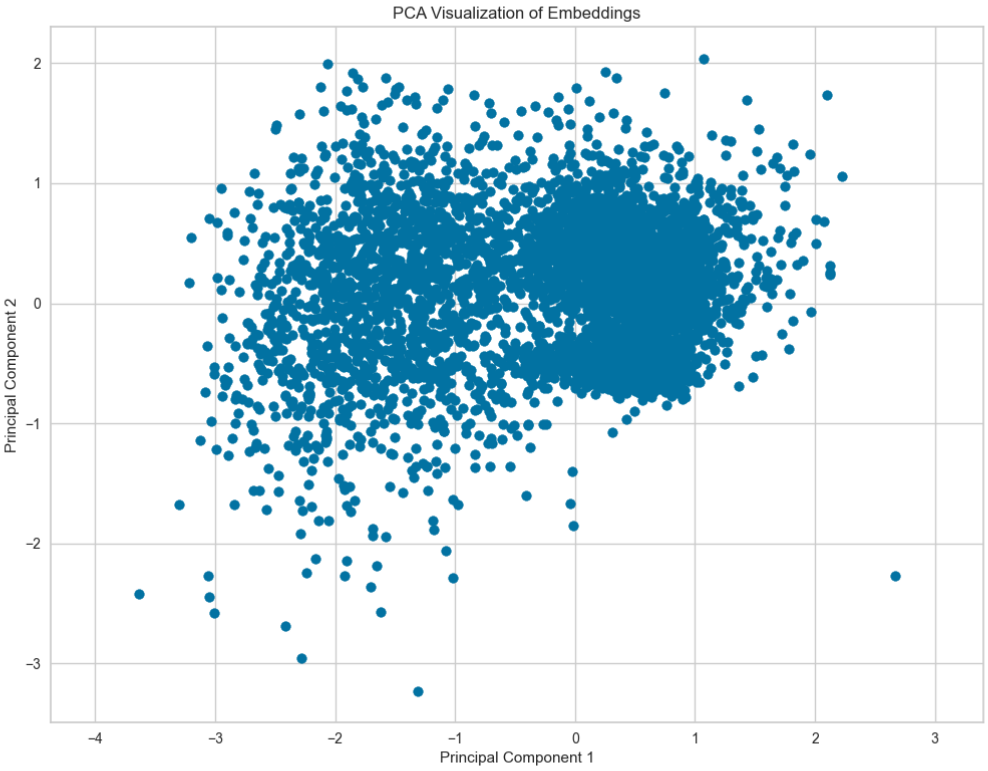


Figure 2: The scatter plot illustrates the distribution of embeddings, originally represented in a 100-dimensional space, reduced to two dimensions using Principal Component Analysis. Each point on the plot corresponds to an individual embedding projected onto the two-dimensional plane defined by the first and second principal components.

- we exclude from the subgraphs all walks on predicates irrelevant to the comparison, by providing a list of stop-patterns (Appendix A)
- we filter out explicitly defined uninformative triples (Appendix B) from the LCS, transforming it into a *Common Subsumer (CS)* that consequently retains only relevant information.

To provide a general understanding of the dataset commonalities, we first compute the CS to the entire dataset, identifying information shared across all resources. This CS is a rather uninformative KG, that is shown in Appendix C and models only one commonality: the fact that all drugs have a type which is *small molecule*[5].

Subsequently, to explain individual clusters, LISE computes the CS for each cluster. A sample CS of Cluster no.58, which contains 52 elements, is shown in Appendix D[6]. In this case, the CS models more commonalities among the 52 cluster items. First, also drugs in Cluster no.58 are of kind small molecule; second, they share a target protein, labeled as `Tyrosine-protein phosphatase non-receptor type 1 [drugbank:BE0000623]`, which is fully described in the CS. In particular, the CS shows that this target affects a human organism type.

Although the generated explanations are already filtered for irrelevant details, LISE further refines them by applying two irrelevance definitions taken from previous work[12]. Specifically, LISE prunes:

- information "irrelevant to the context", *i.e.*, the features shared by the entire dataset, because they

[5] Figure 6 shows this specific CS, verbalized in natural language using the template-based approach described in Section 2.4.
[6] A verbalization of Cluster no.58, using the template-based approach, is shown in Figure 5.

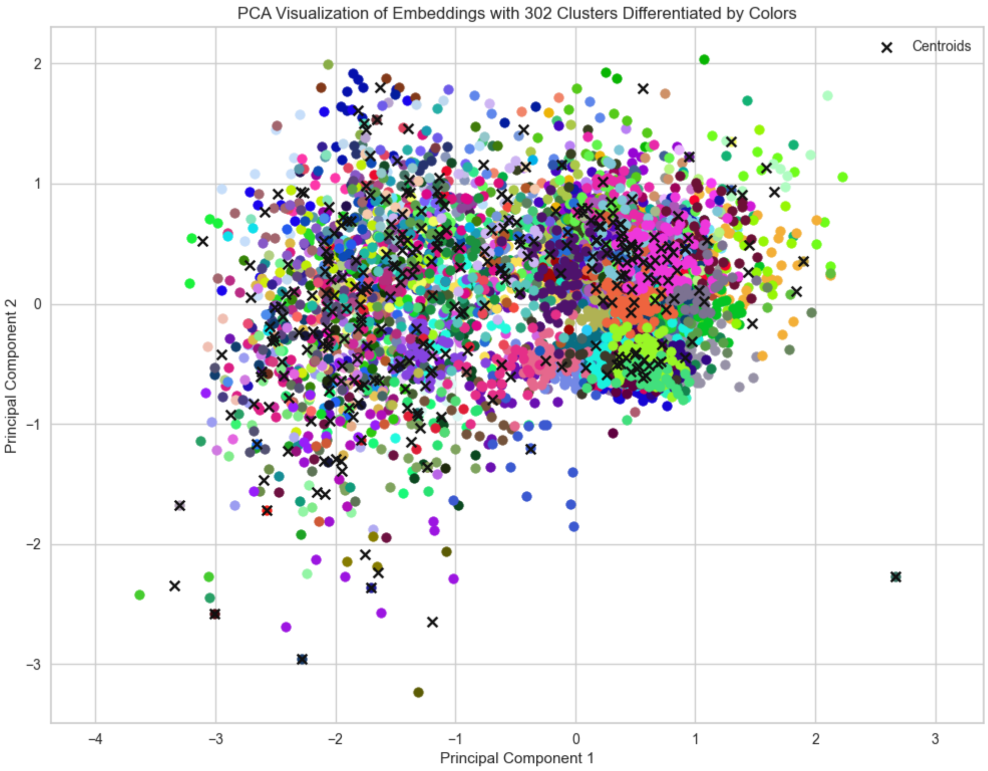


Figure 3: The scatter plot visualizes the distribution of embeddings, originally represented in a 100-dimensional space, reduced to two dimensions using Principal Component Analysis. Each point, colored differently to indicate membership in one of the 302 clusters, represents an embedding projected onto the two-dimensional plane defined by the first two principal components. The black ’x’ markers denote the centroids of the clusters.

do not discriminate clusters.

- information "irrelevant to the user", *i.e.*, information already known to the user, based on his/her Personal KG.

In the use case addressed, information irrelevant to the context is the one listed in Appendix C, stating that all drugs have type *small molecule*.

Regarding irrelevance to the user, in the use case we assume that user knows that some drug targets (Pyridoxal kinase, Cyclin-dependent kinase 2, Glutathione S-transferase A1, Tyrosine-protein kinase JAK2, Tyrosine-protein phosphatase non-receptor type 1) act on human organism. Thus, we model his/her Personal Knowlwedge Graph in RDF, as shown in Appendix E.

By pruning the CS of Cluster no.58 of both types of irrelevant information, we obtain a new CS, shown in Appendix F[7]. This new CS does not include the information that all cluster items have small molecules (irrelevant to the context) and that protein "Tyrosine-protein phosphatase non-receptor type 1" acts on the human organism (already known to the user).

2.4. Natural Language Generation Module

This component is responsible for translating the relevant commonalities modeled by the Common Subsumer, as previously computed and refined by the logic-based component (Section 2.3) of LISE, into human-readable explanations.

[7] The new CS is also shown in Figure 7 in a human-readable format, by using the template-based verbalization approach.

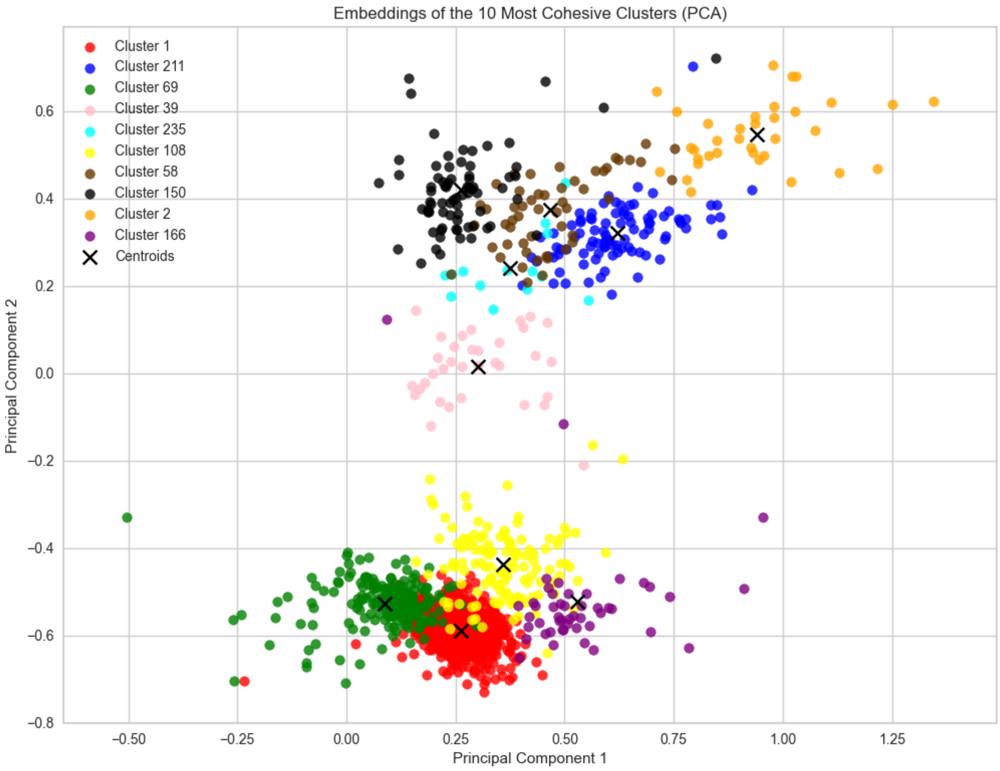


Figure 4: The scatter plot displays the distribution of embeddings from the 10 most cohesive clusters, originally represented in a 100-dimensional space and reduced to two dimensions using Principal Component Analysis. Each colored point represents a data point belonging to one of the selected clusters, while black 'x' markers indicate the centroids of these clusters.

To this end, LISE integrates a previously developed template-based Natural Language Generation tool [13, 14].

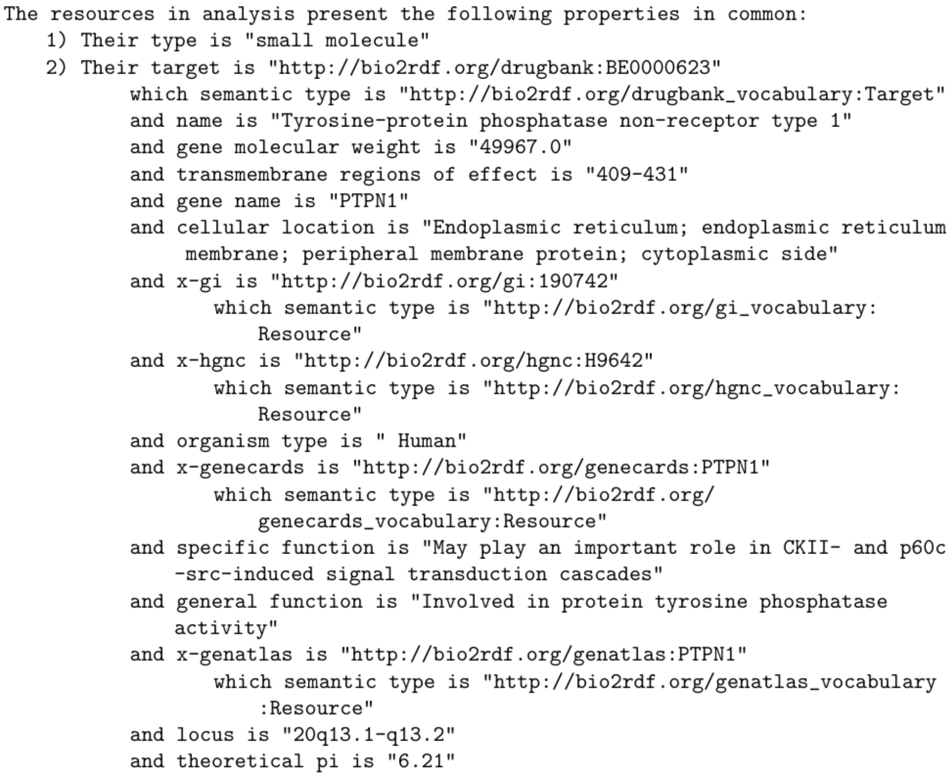
In our use case, LISE produces the explanation in Figure 5 for Cluster no.58, before pruning irrelevance. By pruning information irrelevant both to the context (explained in Figure 6, thanks to the tool) and to the user, the explanation in Figure 7 is generated.

While effective in producing clear and understandable explanations in specific domains, the NLG tool requires the manual definition of a context-specific dictionary, a task that is challenging to automate. To address this limitation, we explored the use of Large Language Models as an alternative to the template-based tool. Specifically, we achieved promising results by training Google Gemini[8] to process RDF Knowledge Graphs, focusing on CSs, which represent the common characteristics to verbalize in natural language. With reference to our use case, we achieved the explanation shown in Figure 8 for Cluster no.58.

Our experiment utilizes the Google Gemini 1.5-Flash[9] model. Results are generated via an API call in Python, which processes the RDF triples in NT format related to a Common Subsumer. The API call, executed with the request "*Verbalize with discursive phrases these RDF NT triples*", is defined with the following system instruction:

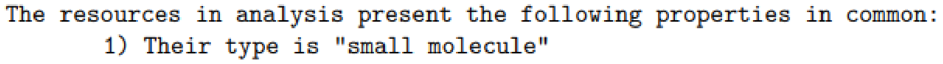
[8] https://ai.google.dev/

[9] https://ai.google.dev/gemini-api/docs/models/gemini?hl=it#gemini-1.5-flash



```
The resources in analysis present the following properties in common:
   1) Their type is "small molecule"
   2) Their target is "http://bio2rdf.org/drugbank:BE0000623"
         which semantic type is "http://bio2rdf.org/drugbank_vocabulary:Target"
         and name is "Tyrosine-protein phosphatase non-receptor type 1"
         and gene molecular weight is "49967.0"
         and transmembrane regions of effect is "409-431"
         and gene name is "PTPN1"
         and cellular location is "Endoplasmic reticulum; endoplasmic reticulum
             membrane; peripheral membrane protein; cytoplasmic side"
         and x-gi is "http://bio2rdf.org/gi:190742"
               which semantic type is "http://bio2rdf.org/gi_vocabulary:
                   Resource"
         and x-hgnc is "http://bio2rdf.org/hgnc:H9642"
               which semantic type is "http://bio2rdf.org/hgnc_vocabulary:
                   Resource"
         and organism type is " Human"
         and x-genecards is "http://bio2rdf.org/genecards:PTPN1"
               which semantic type is "http://bio2rdf.org/
                   genecards_vocabulary:Resource"
         and specific function is "May play an important role in CKII- and p60c
            -src-induced signal transduction cascades"
         and general function is "Involved in protein tyrosine phosphatase
            activity"
         and x-genatlas is "http://bio2rdf.org/genatlas:PTPN1"
               which semantic type is "http://bio2rdf.org/genatlas_vocabulary
                   :Resource"
         and locus is "20q13.1-q13.2"
         and theoretical pi is "6.21"
```

Figure 5: Template-based verbalization of the Common Subsumer obtained from Cluster no.58.



```
The resources in analysis present the following properties in common:
        1) Their type is "small molecule"
```

Figure 6: Template-based verbalization of the Common Subsumer of the entire dataset. Such an information is common to *all* clusters, hence it is irrelevant to the context of describing a single cluster.

```
For each blank node (identified with 'genid') in the text you are given, you must associate a generic variable. Create a dictionary composed of blank nodes with associated variable. Then verbalize in natural language the triples present in the text considering the defined vocabulary and when the variables are repeated you must refer to that one by continuing the same sentence. Starting the sentence with the verbalized form of root node, create discursive sentences.
Verbalize in natural language the URIs content.
Verbalize in natural language the LITERALS content.
Handle blank nodes with general terms.
Return only verbalization.
```

Additionally, the root node of the CS is passed to the API call as a variable. Table 2 shows the parameters of API call.

Temperature	1
Top_p	0.95
Top_k	40
Max_output_tokens	8192
Response_mime_type	"text/plain"

Table 2
Parameter employed in the API call to Gemini for the explanation of Common Sumbumers in RDF.

```
The resources in analysis present the following properties in common:
       1) Their target is "http://bio2rdf.org/drugbank:BE0000623"
              which theoretical pi is "6.21"
              and locus is "20q13.1-q13.2"
              and x-gi is "http://bio2rdf.org/gi:190742"
                     which semantic type is "http://bio2rdf.org/gi_vocabulary:
                         Resource"
              and name is "Tyrosine-protein phosphatase non-receptor type 1"
              and gene molecular weight is "49967.0"
              and x-hgnc is "http://bio2rdf.org/hgnc:H9642"
                     which semantic type is "http://bio2rdf.org/hgnc_vocabulary:
                         Resource"
              and transmembrane regions of effect is "409-431"
              and x-genatlas is "http://bio2rdf.org/genatlas:PTPN1"
                     which semantic type is "http://bio2rdf.org/genatlas_vocabulary
                         :Resource"
              and general function is "Involved in protein tyrosine phosphatase
                  activity"
              and specific function is "May play an important role in CKII- and p60c
                  -src-induced signal transduction cascades"
              and x-genecards is "http://bio2rdf.org/genecards:PTPN1"
                     which semantic type is "http://bio2rdf.org/
                         genecards_vocabulary:Resource"
              and gene name is "PTPN1"
              and semantic type is "http://bio2rdf.org/drugbank_vocabulary:Target"
              and cellular location is "Endoplasmic reticulum; endoplasmic reticulum
                   membrane; peripheral membrane protein; cytoplasmic side"
```

Figure 7: Template-based verbalization of the Common Subsumer obtained from Cluster no.58 pruned of irrelevant information.

A comparison between the explanations generated by the template-based tool (Figure 5) and Google Gemini (Figure 8) reveals comparability in terms of clarity and interpretability.

```
The entity is a small molecule and its target is Tyrosine-protein phosphatase
    non-receptor type 1, which is involved in protein tyrosine phosphatase
    activity and may play an important role in CKII- and p60c-src-induced signal
    transduction cascades. This target has a molecular weight of 49967.0, a
    theoretical pI of 6.21, and is located in the endoplasmic reticulum,
    endoplasmic reticulum membrane, and cytoplasm. Its gene name is PTPN1,
    located at 20q13.1-q13.2, and has GI number 190742. Additionally, it has
    409-431 transmembrane regions and is found in humans. The entity has gene
    cards, HGNc, and GenAtlas identifiers associated with it. The target is also
    a resource and is related to other resources, including GeneCards, HGNc and
    GenAtlas resources.
```

Figure 8: Gemini verbalization of the Common Subsumer obtained from Cluster no.58.

Although both approaches require a degree of customization, Gemini's APIs enable semi-automated training, offering greater flexibility and adaptability to different domains. Despite these advantages, LISE continues to use the template-based tool, primarily due to system modularity. In fact, the output of the NLG module is processed and passed to the User Interaction and Feedback Loop module, which collects user feedback on the relevance of the generated explanations. Currently, these module requires explanations to follow the template structure, limiting the immediate adoption of LLM-based solutions. We are therefore investigating strategies to make the interactive components independent of the NLG template, thereby increasing system flexibility.

2.5. User Interaction and Feedback Loop

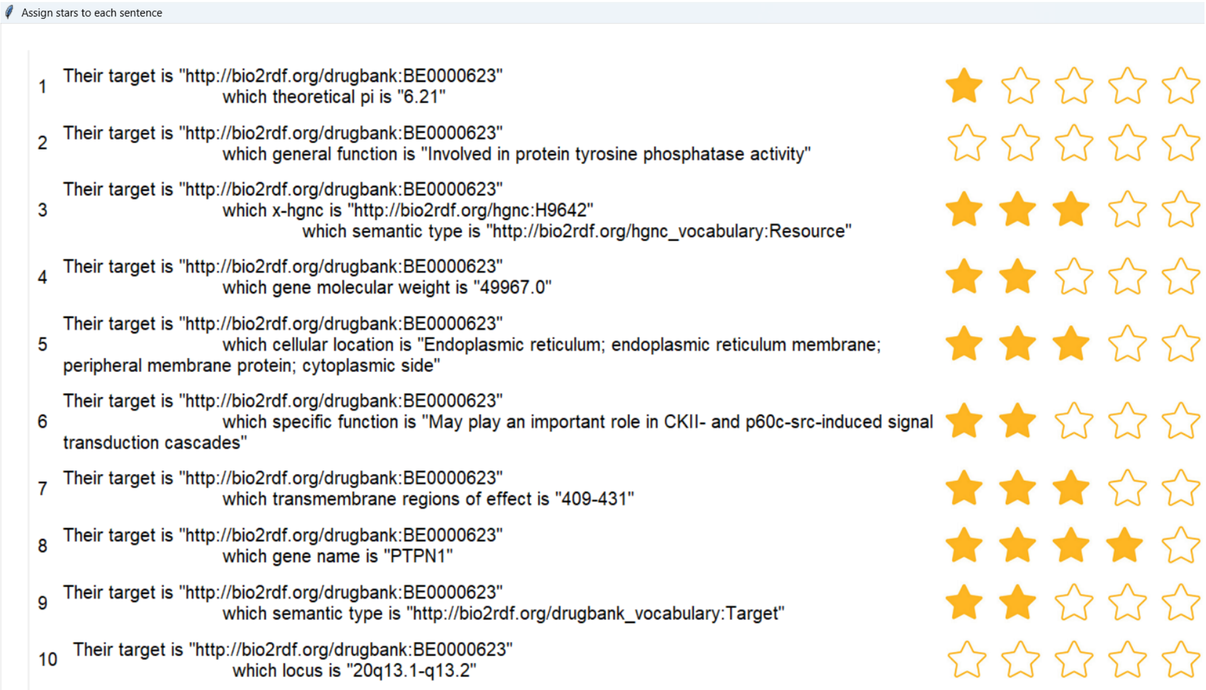
In this section, we present the human-in-the-loop interactive approach used to let the user evaluate and refine explanations generated by the Natural Language Generation component (Section 2.4).

LISE collects user feedback through a Graphical User Interface (GUI) developed using the Tkinter[10]

[10] https://docs.python.org/3/library/tkinter.html

Python library. This interface enables users to rate explanation sentences derived from a logically computed CS that abstracts cluster commonalities. The collected ratings are subsequently leveraged to enhance the system's ability to predict relevance scores for new explanations.

Figure 9 shows the GUI corresponding to the CS of Cluster no.58, pruned of irrelevant information.



1 Their target is "http://bio2rdf.org/drugbank:BE0000623" which theoretical pi is "6.21" ★☆☆☆☆

2 Their target is "http://bio2rdf.org/drugbank:BE0000623" which general function is "Involved in protein tyrosine phosphatase activity" ☆☆☆☆☆

3 Their target is "http://bio2rdf.org/drugbank:BE0000623" which x-hgnc is "http://bio2rdf.org/hgnc:H9642" which semantic type is "http://bio2rdf.org/hgnc_vocabulary:Resource" ★★★☆☆

4 Their target is "http://bio2rdf.org/drugbank:BE0000623" which gene molecular weight is "49967.0" ★★☆☆☆

5 Their target is "http://bio2rdf.org/drugbank:BE0000623" which cellular location is "Endoplasmic reticulum; endoplasmic reticulum membrane; peripheral membrane protein; cytoplasmic side" ★★★☆☆

6 Their target is "http://bio2rdf.org/drugbank:BE0000623" which specific function is "May play an important role in CKII- and p60c-src-induced signal transduction cascades" ★★☆☆☆

7 Their target is "http://bio2rdf.org/drugbank:BE0000623" which transmembrane regions of effect is "409-431" ★★★☆☆

8 Their target is "http://bio2rdf.org/drugbank:BE0000623" which gene name is "PTPN1" ★★★★☆

9 Their target is "http://bio2rdf.org/drugbank:BE0000623" which semantic type is "http://bio2rdf.org/drugbank_vocabulary:Target" ★★☆☆☆

10 Their target is "http://bio2rdf.org/drugbank:BE0000623" which locus is "20q13.1-q13.2" ☆☆☆☆☆

Figure 9: Star-rating user feedback: screenshot of LISE graphical user interface to allow users to evaluate the relevance of sentences extracted from the generated explanation in the Logic-based Module. The template-based verbalization ensures that each evaluation is one-one with an RDF pattern in the Common Subsumer.

The interface presents the generated explanation in multiple sentences, each associated with an RDF pattern in the Common Subsumer. Users provide feedback via a star-rating system ranging from 1 to 5 stars. These ratings are subsequently normalized into numerical values on a scale from 0 (0 stars) to 1 (5 stars) in increments of 0.2.

The user-provided ratings are stored in a structured data model, where each RDF pattern verbalized in an explanation sentence is assigned a weight. Once feedback is collected, it is used as a dataset for training a linear regression model to predict relevance scores for RDF patterns. In the current version, LISE implements the 'LinearRegression' model from the Scikit-learn Python library [10].

The weights estimated by the regression model are then used to refine the embedding generation process. In particular, the *Predicate Relevance Weight* sampling strategy we proposed in Section 2.2 adopts weights learned by the regression model. Consequently, by training RDF2VeC, LISE follows a relevance-focused sampling strategy that, while computing the embeddings, gives priority to the patterns more relevant to users.

3. Conclusion and Future Work

LISE introduces an approach for explaining the clustering of RDF resources by combining machine learning, knowledge-based reasoning, and natural language generation within an interactive feedback loop. Its characteristics of being text-based, method-agnostic, and post-hoc ensure flexibility and adaptability, making LISE a valuable tool to enhance the interpretability of clustering models. The results obtained from an RDF dataset containing structured drug-related information demonstrate the effectiveness in generating comprehensible and relevant explanations for users. The integration of human feedback enables the progressive refinement of vector representations, improving the alignment between the generated explanations and user expectations. However, the current template-based natural language generation method presents limitations in terms of flexibility. For this reason, as part of our future work, we plan to replace the template-based approach with one employing Large Language Models (but still allowing the user to choose what information is considered relevant) to overcome these constraints and further enhance the quality of generated explanations. Moreover, we intend to explore the use of different Knowledge Graph embeddings models to more accurately capture semantic similarities between cluster resources.

Acknowledgments

We acknowledge partial support by the italian aerospace technological district "Distretto Tecnologico Aerospaziale S.C.a.r.l." and project "LIFE: the itaLian system wIde Frailty nEtwork" founded by Ministry of Health (CUP D93C22000640001).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, G. Falkman, Interactive clustering: A comprehensive review, ACM Computing Surveys (CSUR) 53 (2020) 1–39.

[2] R. T. Sousa, S. Silva, C. Pesquita, Supervised biomedical semantic similarity, IEEE Access 11 (2023) 60635–60645. doi:10.1109/ACCESS.2023.3285406.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[4] P. Ristoski, J. Rosati, T. D. Noia, R. D. Leone, H. Paulheim, Rdf2vec: Rdf graph embeddings and their applications, Semantic Web 10 (2019) 721–752. URL: https://madoc.bib.uni-mannheim.de/50498/. doi:10.3233/SW-180317.

[5] S. Colucci, F. M. Donini, E. D. Sciascio, Logical comparison over RDF resources in bio-informatics, J. Biomed. Informatics 76 (2017) 87–101. URL: https://doi.org/10.1016/j.jbi.2017.11.004. doi:10.1016/J.JBI.2017.11.004.

[6] B. Steenwinckel, G. Vandewiele, P. Bonte, M. Weyns, H. Paulheim, P. Ristoski, F. De Turck, F. Ongenae, Walk extraction strategies for node embeddings with rdf2vec in knowledge graphs, in: G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, S. Khan (Eds.), Database and Expert Systems Applications - DEXA 2021 Workshops, Springer International Publishing, Cham, 2021, pp. 70–80.

[7] M. Cochez, P. Ristoski, S. P. Ponzetto, H. Paulheim, Biased graph walks for rdf graph embeddings, in: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS '17, Association for Computing Machinery, New York, NY, USA, 2017. URL: https://doi.org/10.1145/3102254.3102279. doi:10.1145/3102254.3102279.

[8] N. Jain, J. Kalo, W. Balke, R. Krestel, Do embeddings actually capture knowledge graph semantics?, in: R. Verborgh, K. Hose, H. Paulheim, P. Champin, M. Maleshkova, Ó. Corcho, P. Ristoski, M. Alam (Eds.), The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings, volume 12731 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 143–159. URL: https://doi.org/10.1007/978-3-030-77385-4_9. doi:10.1007/978-3-030-77385-4_9.

[9] N. Hubert, H. Paulheim, A. Brun, D. Monticolo, Do similar entities have similar embeddings?, in: A. Meroño Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, P. Lisena (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2024, pp. 3–21.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[11] S. Colucci, F. M. Donini, E. Di Sciascio, Computing the commonalities of clusters in resource description framework: Computational aspects, Data 9 (2024). URL: https://www.mdpi.com/2306-5729/9/10/121. doi:10.3390/data9100121.

[12] S. Colucci, F. M. Donini, E. Di Sciascio, On the relevance of explanation for RDF resources similarity, in: Model-Driven Organizational and Business Agility - Third International Workshop, MOBA 2023, volume 488 of *LNBIP*, Springer, 2023, pp. 96–107.

[13] S. Colucci, F. M. Donini, N. Iurilli, E. D. Sciascio, A business intelligence tool for explaining similarity, in: E. Babkin, J. Barjis, P. Malyzhenkov, V. Merunka (Eds.), Model-Driven Organizational and Business Agility - Second International Workshop, MOBA 2022, Leuven, Belgium, June 6-7, 2022, Revised Selected Papers, volume 457 of *Lecture Notes in Business Information Processing*, Springer, 2022, pp. 50–64. URL: https://doi.org/10.1007/978-3-031-17728-6_5. doi:10.1007/978-3-031-17728-6_5.

[14] S. Colucci, F. M. Donini, E. D. Sciascio, Explaining commonalities of clusters of RDF resources in natural language, in: A. Appice, H. Azzag, M. Hacid, A. Hadjali, Z. W. Ras (Eds.), Foundations of Intelligent Systems - 27th International Symposium, ISMIS 2024, Poitiers, France, June 17-19, 2024, volume 14670 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 160–169. URL: https://doi.org/10.1007/978-3-031-62700-2_15. doi:10.1007/978-3-031-62700-2_15.

[15] D. Beckett, T. Berners-Lee, Turtle - Terse RDF Triple Language, W3C Team Submission, 2011. URL: http://www.w3.org/TeamSubmission/turtle/.

A. Stop Patterns

In the appendices, we use the RDF Turtle syntax [15], and refer to the following prefixes:

```
@prefix drugbank_vocabulary: <http://bio2rdf.org/drugbank_vocabulary:>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix owl: <http://www.w3.org/2002/07/owl#>
@prefix bio2rdf_vocabulary: <http://bio2rdf.org/bio2rdf_vocabulary:>
@prefix dct: <http://purl.org/dc/terms/>
@prefix void: <http://rdfs.org/ns/void#>
@prefix bio2rdf: <http://bio2rdf.org/>
```

Stop-patterns (determined heuristically) are patterns of RDF triples considered irrelevant for explaining similarity. They are constituted by all triples <*s p o*> meeting at least one of the following criteria:

- $p \in \{dct : description,$
  $dct : identifier,$
  $dct : title,$
  $rdfs : label,$
  $rdfs : seeAlso,$
  $owl : sameAs,$
  $void : inDataset,$
  $bio2rdf\_vocabulary : identifier,$

$bio2rdf\_vocabulary : namespace,$
$bio2rdf\_vocabulary : uri,$
$bio2rdf\_vocabulary : x - identifiers.org,$
$drugbank\_vocabulary : id,$
$drugbank\_vocabulary : brand,$
$drugbank\_vocabulary : carrier,$
$drugbank\_vocabulary : clearance,$
$drugbank\_vocabulary : dosage,$
$drugbank\_vocabulary : group,$
$drugbank\_vocabulary : indication,$
$drugbank\_vocabulary : manufacturer,$
$drugbank\_vocabulary : patent,$
$drugbank\_vocabulary : product,$
$drugbank\_vocabulary : synonym,$
$drugbank\_vocabulary : transporter,$
$drugbank\_vocabulary : volume - of - distribution,$
$drugbank\_vocabulary : x - wikipedia,$
$drugbank\_vocabulary : x - ahfs,$
$drugbank\_vocabulary : x - atc,$
$drugbank\_vocabulary : x - bindingdb,$
$drugbank\_vocabulary : x - chebi,$
$drugbank\_vocabulary : x - chemspider,$
$drugbank\_vocabulary : x - genbank,$
$drugbank\_vocabulary : x - gtp,$
$drugbank\_vocabulary : x - iuphar,$
$drugbank\_vocabulary : x - kegg,$
$drugbank\_vocabulary : x - ndc,$
$drugbank\_vocabulary : x - pdb,$
$drugbank\_vocabulary : x - pharmgkb,$
$drugbank\_vocabulary : x - pubchemcompound,$
$drugbank\_vocabulary : x - pubchemsubstance,$
$drugbank\_vocabulary : x - uniprot,$
$drugbank\_vocabulary : calculated - properties,$
$drugbank\_vocabulary : experimental - properties\}$

- $p = rdf : type$ and $o \in$
  $\{owl : Class,$
  $owl : ObjectProperty,$
  $owl : DatatypeProperty,$
  $drugbank\_vocabulary : Resource,$
  $drugbank\_vocabulary : Substructures,$
  $drugbank\_vocabulary : Drug\}$

B. Uninformative Triples

Uninformative triples are those <*s p o*> triples meeting at least one of the conditions listed below, provided that *o* has no successors:

- $p \in \{rdf : type,$
  $rdfs : seeAlso,$
  $rdfs : subClassOf,$
  $rdfs : label,$

$drugbank\_vocabulary : type,$
$rdf : value,$
$drugbank\_vocabulary : substructure,$
$drugbank\_vocabulary : name,$
$drugbank\_vocabulary : organism\}$

- p is a blank node

C. Common Subsumer of the entire dataset

The following triple represents the CS of the entire dataset of drugs selected from Drugbank:

```
_:N469a9065937b4f73a25076fe400b2502 drugbank_vocabulary:type drugbank_vocabulary:Small-molecule .
```

We recall that all identifiers starting with _: denote a blank node.

D. Common Subsumer of Cluster no.58

In what follows, the complete CS of cluster no.58 is shown. The CS is an RDF knowledge graph rooted in the blank node `_:Nc9029d61afbb426e8559468beee0277f`.

```
_:Nc9029d61afbb426e8559468beee0277f drugbank_vocabulary:target bio2rdf:drugbank:BE0000623 ;
    drugbank_vocabulary:type drugbank_vocabulary:Small-molecule .

bio2rdf:drugbank:BE0000623 a drugbank_vocabulary:Target ;
    drugbank_vocabulary:cellular-location "Endoplasmic reticulum;
    endoplasmic reticulum membrane; peripheral membrane protein; cytoplasmic side" ;
    drugbank_vocabulary:gene-name "PTPN1" ;
   drugbank_vocabulary:general-function "Involved in protein tyrosine phosphatase activity" ;
    drugbank_vocabulary:locus "20q13.1-q13.2" ;
    drugbank_vocabulary:molecular-weight "49967.0" ;
    drugbank_vocabulary:name "Tyrosine-protein phosphatase non-receptor type 1" ;
    drugbank_vocabulary:organism "Human" ;
    drugbank_vocabulary:specific-function "May play an important role in CKII- and
    p60c-src-induced signal transduction cascades" ;
    drugbank_vocabulary:theoretical-pi "6.21" ;
    drugbank_vocabulary:transmembrane-regions "409-431" ;
    drugbank_vocabulary:x-genatlas bio2rdf:genatlas:PTPN1 ;
    drugbank_vocabulary:x-genecards bio2rdf:genecards:PTPN1 ;
    drugbank_vocabulary:x-gi bio2rdf:gi:190742 ;
    drugbank_vocabulary:x-hgnc bio2rdf:hgnc:H9642 .

bio2rdf:genatlas:PTPN1 a bio2rdf:genatlas_vocabulary:Resource .

bio2rdf:genecards:PTPN1 a bio2rdf:genecards_vocabulary:Resource .

bio2rdf:gi:190742 a bio2rdf:gi_vocabulary:Resource .

bio2rdf:hgnc:H9642 a bio2rdf:hgnc_vocabulary:Resource .
```

E. Personal Knowledge Graph

The following triples set represents the Personal Knowledge Graph of an hypothetical user of LISE, used in the addressed use case:

```
bio2rdf:drugbank:BE0002281 drugbank_vocabulary:organism "Human" .
bio2rdf:drugbank:BE0000042 drugbank_vocabulary:organism "Human" .
bio2rdf:drugbank:BE0001072 drugbank_vocabulary:organism "Human" .
bio2rdf:drugbank:BE0002408 drugbank_vocabulary:organism "Human" .
bio2rdf:drugbank:BE0000623 drugbank_vocabulary:organism "Human" .
```

F. Common Subsumer of Cluster no.58 pruned of irrelevant information

In what follows, it is shown the CS of cluster no.58 pruned of irrelevant information. The CS is an RDF knowledge graph rooted in the blank node `_:Nc9029d61afbb426e8559468beee0277f`.

```
_:Nc9029d61afbb426e8559468beee0277f drugbank_vocabulary:target bio2rdf:drugbank:BE0000623 .

bio2rdf:drugbank:BE0000623 a drugbank_vocabulary:Target ;
    drugbank_vocabulary:cellular-location "Endoplasmic reticulum; endoplasmic reticulum
    membrane; peripheral membrane protein; cytoplasmic side" ;
    drugbank_vocabulary:gene-name "PTPN1" ;
    drugbank_vocabulary:general-function "Involved in protein tyrosine phosphatase activity" ;
    drugbank_vocabulary:locus "20q13.1-q13.2" ;
    drugbank_vocabulary:molecular-weight "49967.0" ;
    drugbank_vocabulary:name "Tyrosine-protein phosphatase non-receptor type 1" ;
   drugbank_vocabulary:specific-function "May play an important role in CKII- and p60c-src-induced
    signal transduction cascades" ;
    drugbank_vocabulary:theoretical-pi "6.21" ;
    drugbank_vocabulary:transmembrane-regions "409-431" ;
    drugbank_vocabulary:x-genatlas bio2rdf:genatlas:PTPN1 ;
    drugbank_vocabulary:x-genecards bio2rdf:genecards:PTPN1 ;
    drugbank_vocabulary:x-gi bio2rdf:gi:190742 ;
    drugbank_vocabulary:x-hgnc bio2rdf:hgnc:H9642 .

bio2rdf:genatlas:PTPN1 a bio2rdf:genatlas_vocabulary:Resource .

bio2rdf:genecards:PTPN1 a bio2rdf:genecards_vocabulary:Resource .

bio2rdf:gi:190742 a bio2rdf:gi_vocabulary:Resource .

bio2rdf:hgnc:H9642 a bio2rdf:hgnc_vocabulary:Resource .
```