Scientific Visualization, 2025, volume 17, number 5, pages 83 - 93, DOI: 10.26583/sv.17.5.09

Data Classification with Using Visualization Tools

Andrey Dzengelewski[1]

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russia

[1] ORCID: 0009-0002-4770-1421, Dzengelewski@gmail.com

Abstract

This article discusses ways to use visualization tools to build object classifiers during automation of a large enterprise. The proposed approaches allow stakeholders to get a visual representation and participate in the decisions required when building a classifier for large arrays of records.

The use of visualization tools is considered when selecting classification objects, determining the attributes and values of classification attributes, ensuring the convenience of the classifier and implementing conflicting requirements from stakeholders. Among the proposed solutions, the methods of using system classes, building logical and physical models of the classifier, multidimensional classification, attribute-value data model, logical data model for describing the required analytics are described.

The subject area is a classifier of works and services, examples of using the proposed solutions and the results of building a classifier at a large enterprise are given.

Keywords: classification, stakeholder, view, class attribute, logical and physical levels of classification, multidimensional classification, attribute-value model, conceptual data model, logical data model.

1. Introduction

The construction of "correct classifiers has always played and continues to play an important role in the automation of enterprises, which was demonstrated at the 2023 conference “The Role of Standardization in the Digitalization of Industry” [1]. For a large enterprise, the most important classification objects are Resources, Works/services and Assets.

When constructing an ontological model of an enterprise, classification plays an important role for constructing first a taxonomy and then an ontology of the enterprise as a whole, which was demonstrated at international ontology conferences [2].

However, correct classification plays no less a role in solving practical problems of automation of specific organizations, especially large ones. When implementing an enterprise master data, it is important to build a convenient and effective tool for working with a large set of items. At the same time, an important part of the work is to coordinate decisions made with stakeholders - business users, who are sometimes far from information technology.

According to TOGAF recommendations [3] one of the tasks of an information system architect is to offer the most convenient views for perception by stakeholders with a wide range of interests. In order not to solve such a problem, it is necessary to offer a method for preparing such a viewpoint for each typical case.

2. Existing methods of data visualization

A review of software tools for displaying geometric objects is given in an article by MEPhI scientists [4]. An overview of information visualization tools is given in the article [5] and other works by Bauman Moscow State Technical University employees. A representative col-

lection of data visualization templates is presented by the Ferdio agency as a result of the data visualization project [6]. The article [7] provides a general set of template applications, including for displaying hierarchical classification.

This paper describes the possibilities and results of practical application of visualization tools for constructing a qualitative classification for a large enterprise.

3. Existing approaches to constructing a hierarchical classification

The basic rules of hierarchical classification are given in the monograph [8]. Modern classification experts use approximately the same set of principles on projects. An option taking into account the capabilities of modern systems is given in [9] and also includes requirements of completeness, consistency, and the presence of a classification feature at each level.

4. Existing classification problems

However, in practice, when constructing real classifiers, problems arise that require non-obvious solutions. Here are some of them:

- selection of classification objects;
- selection of classification features;
- ensuring the convenience of the classifier while observing formal rules;
- implementation of conflicting requirements for the composition of classes.

Let's consider options for solving these problems using visualization tools using the example of constructing a classifier of works and services.

5. Methods for solving classification problems using visualization tools

5.1. Selection of classification objects

To correctly select classification objects, it is necessary to proceed from the goals of constructing and further using the classifier and configuration of integrated information systems. For example, detailing of services is usually required when concluding contracts with suppliers or consumers of services, when preparing specifications for equipment repair, less often for capital construction, when accounting for costs. If there are a small number of areas, this can be visualized using a Venn diagram (Figure 1).

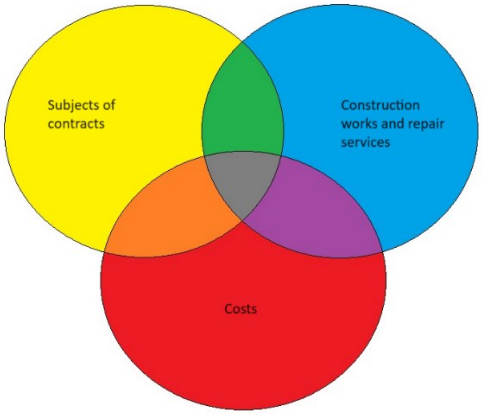

Figure 1. Initial record sets for the corporate directory “Works and services”

If the task is to integrate three systems of the corresponding purpose, then the completeness of the directory can be determined by the areas of intersection. If several systems oper-

ate in each area, then a classifier covering the combination of areas will be required. This solution will allow not only to perform a consolidated analysis of works and services, but also to provide uniform corporate accounting features for the entire enterprise.

5.2. Selection of classification features

To select mandatory classification features, the requirements of the areas of use may be considered. However, there is a risk of including specific attributes inherent only to a given area. For example, in accounting, the activities of contractors can be written off immediately in the reporting period as services, or accumulated on cost objects as work performed. In general, such activities can be both work and services, and such a sign cannot be used.

At the same time, if it is required to generate reports according the content of services and such a sign can be used for classification.

Another difficulty is that feature must be made at each classification level. Here it is useful to look at other directories of the organization for which the general task of classifying all activities has already been solved. Typically, the following classification criteria are used at the top level: divisions, business processes, functional areas, main services and products.

Further, the features will depend on the composition of the classified records. The Section Layer Diagram (Figure 2) or by classification [6] «Icicle» allows you to depict not only the classification branches, but also the values of the features used diagram. The difference is that for our purposes it is expanded by 90 degrees and supplemented with important class - attributes and values of attributes of higher classes.

Level 2 class sign	Meaning sign	Level 2 class	Level 3 class sign	Meaning sign	Level 3 class	Level 4 class sign	Meaning of the sign	Level 4 class
Degree of participation in the registration of transportation	Shipping	**Transportation services**	Object of transportation	People	**Transportation of employees**			
				Documents	**Delivery of documents**			
				Cargo	**Cargo transportation**	**Kind of transport**	**Automobile**	**Transportation by car transport**
							Railway	**Transportation by rail**
							Other	**Transportation by other modes of transport**
	Transportation formalities	**Services not related to transportation**	Type of work	Decor	**Registration services**			
				Re-felling	**Transshipment services**			
				Service service	**Vehicle servicing services**			
	Complete services	**Forwarding services**						

Figure 2. Using a section layer diagram to show features and their meanings

5.3. Ensuring the convenience of the classifier while observing formal rules

One of the common mistakes when building classifiers is the desire to “even out” the number of levels and the length of classification branches at all costs. In this case, the criterion of “convenience” is approximately the same complexity of each level.

At first glance, such a classifier is more convenient to use, since it looks more streamlined visually. However, this is a deceptive impression. With this approach, the principle of identifying a clear classification feature at each level is violated. Complex subject areas are usually not symmetrical. In terms of functional modeling, such an observation was described in the methodology for functional modeling [10], but this rule is also applicable to data classification. Moreover, if you see a “beautiful” symmetrical classification, then it is already visually clear that it is most likely incorrect.

On the other hand, if we strictly follow the rule of one class - one attribute, then in practice we can get a logically correctly constructed, but absolutely inconvenient to use classifier.

The solution here is to divide the construction of a classification tree into two stages: the construction of a logical model and the option of physical model for the final implementation in the system, similar to the approach to data modeling [11]. At the logical level, it is useful to clearly record all possible features and strict division at each level. At the physical level, you can carefully combine the most branched sections, taking into account the knowledge of the previous level, transferring the abbreviated features into the class name.

In Figure 3, when using a standard tree, we see that we can shorten the third classification branch by combining the attributes “Object of transportation” and “Mode of transport”.

5.4. Implementation of conflicting requirements for class composition

In a large enterprise, there may be different requirements for the required classes. For example, in a purchasing department (or holding organization), it is important to take into account the transportation of goods by mode of transport in order to choose the most effective methods. And for a sales department/organization it may be important to divide cargo transportation services by product type for easier cost calculation.

In principle, such a problem is solved by the Cartesian product of the required classes. This approach is acceptable within one organization or for small volumes. However, within the holding, each organization may require its own classification feature at the same generalization.

This limitation also has a solution method within a single hierarchical classifier. It consists of system classification features, where the feature is the sign of the class of the next level. In our example, the feature will be “Classification Feature", and the values will be “Object of transportation” and “Mode of transport” (Figure 4).

Figure 3. Reducing the length of classification branches in the Tree diagram

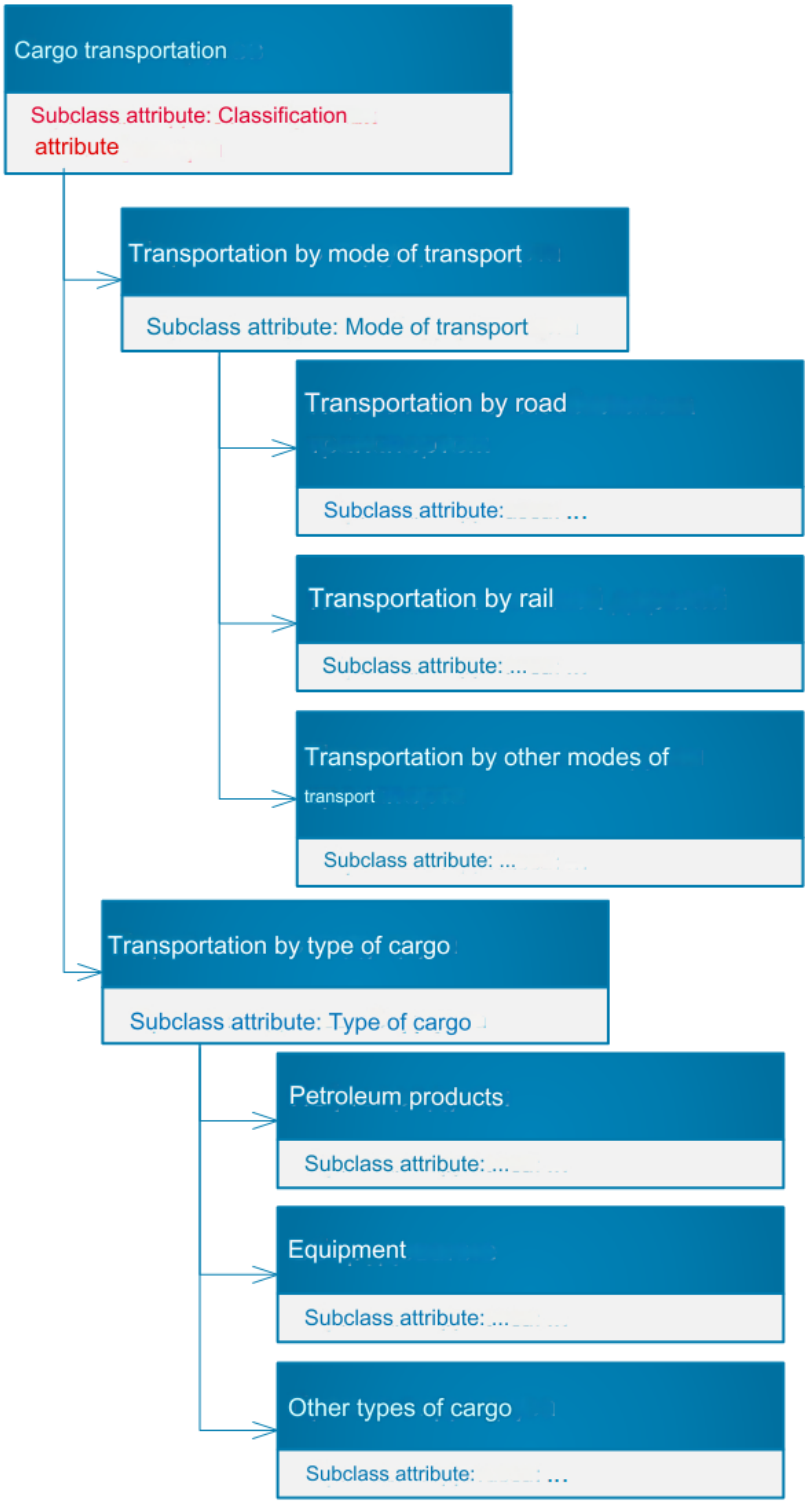

Figure 4. Using the system class feature

It should be noted that this method is only applicable if

1) classes below this level are used only locally; all corporate attributes can be linked at levels.
2) the use of parallel classifications is significantly less effective: it requires additional development in related systems, complicates data models, etc.

A more promising approach is to go beyond the use of a single hierarchy. More complex data models can help to reflect the necessary entities and relationships can help to find the most effective solution options. Three solution methods will be shown below, for the presentation of, which data modeling techniques are used. Among the many possible notations [12], [13] the most convenient notation has proven to be the visualization of relationships using

Martin's notation (another name is “Crow's Foot”), which more clearly reflects the cardinality of the related entities and compactly represents a set of attributes.

3.1) Multidimensional classification

If the same objects must be classified using different classification criteria, then perhaps the option of moving to facet classification is suitable. A conceptual data model showing this solution method is shown in Figure 5.

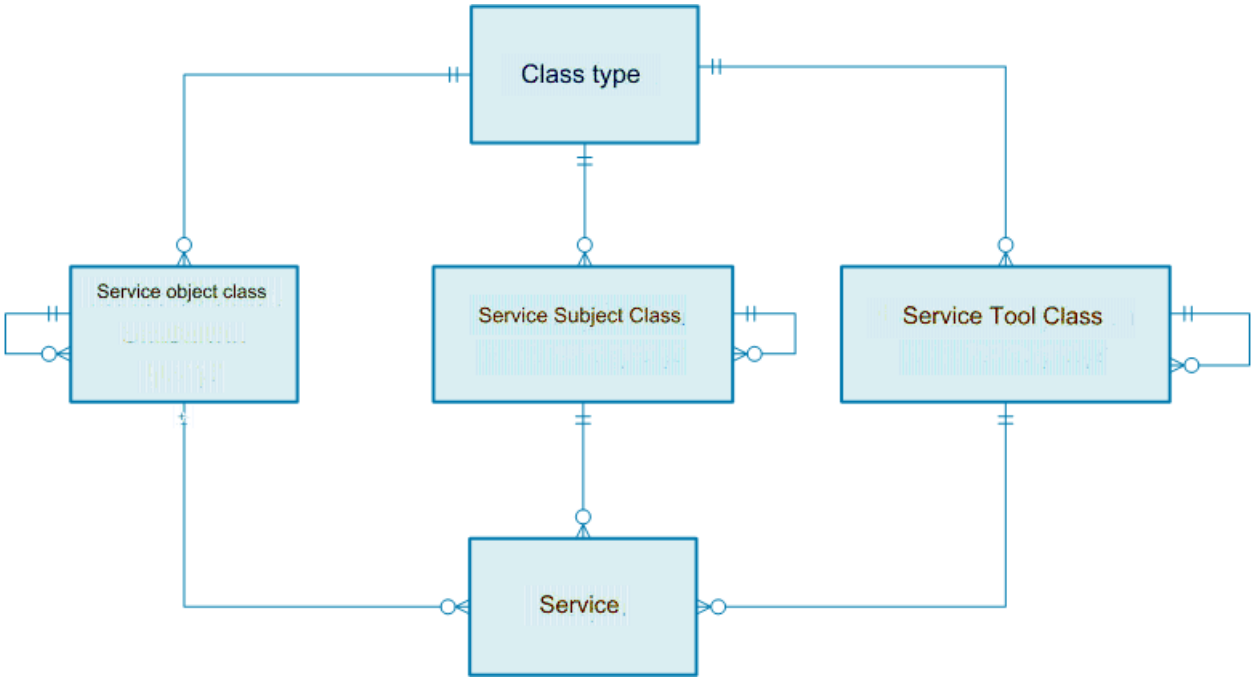

Figure 5. Conceptual data model of multidimensional classification of services

Multidimensional classification opens up additional possibilities for visual data analysis. Figures 5 and 6 show examples of diagrams using multidimensional classification of services.

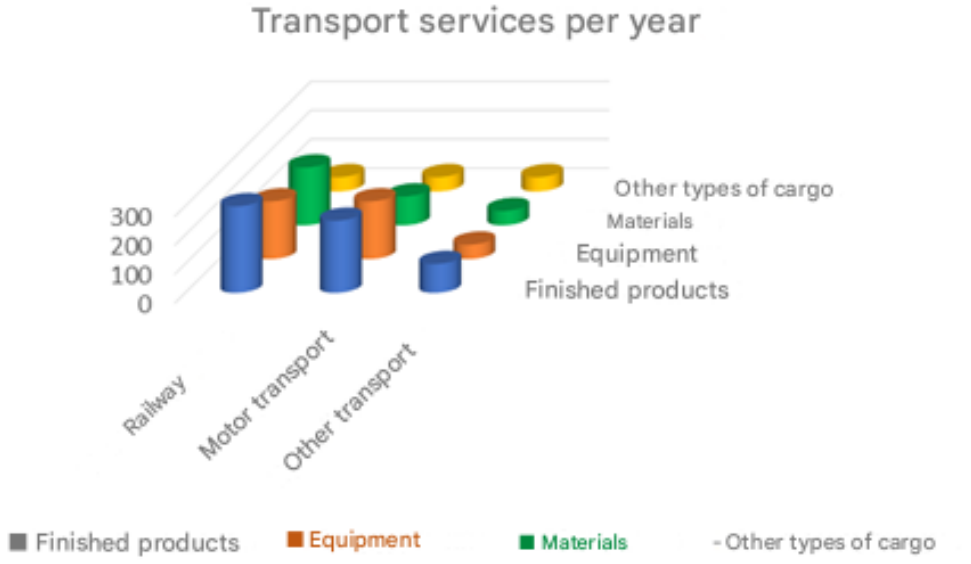

Figure 6. Volume histogram

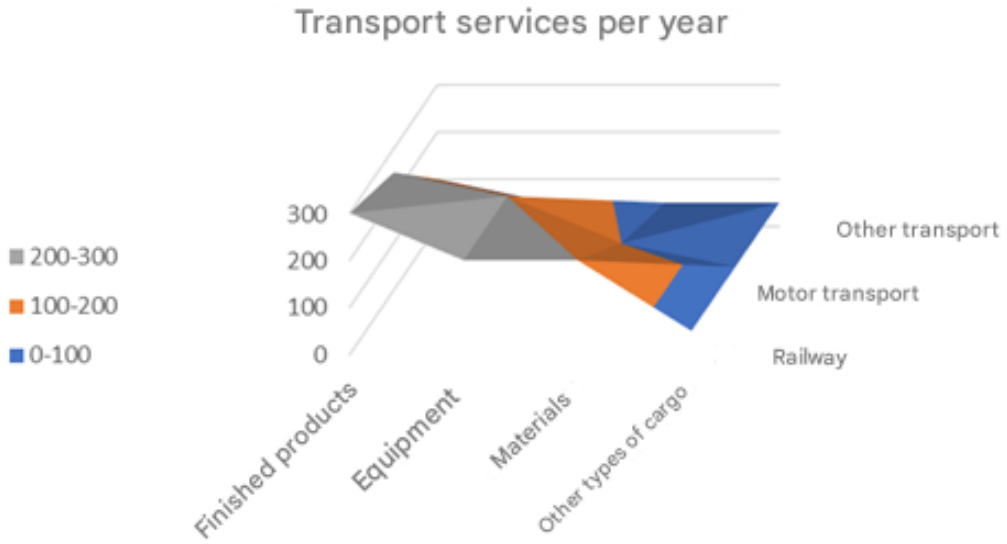

Figure 7. Surface diagram

3.2) If the set of additional features for each subclass is unique and contains a large number of linear values, we can use another approach that involves storing our own sets of features for each class (Figure 6). It should be noted that this approach is most effective for classifying materials. For more information on solving problems of automating the directory of materials, see [14].[15]

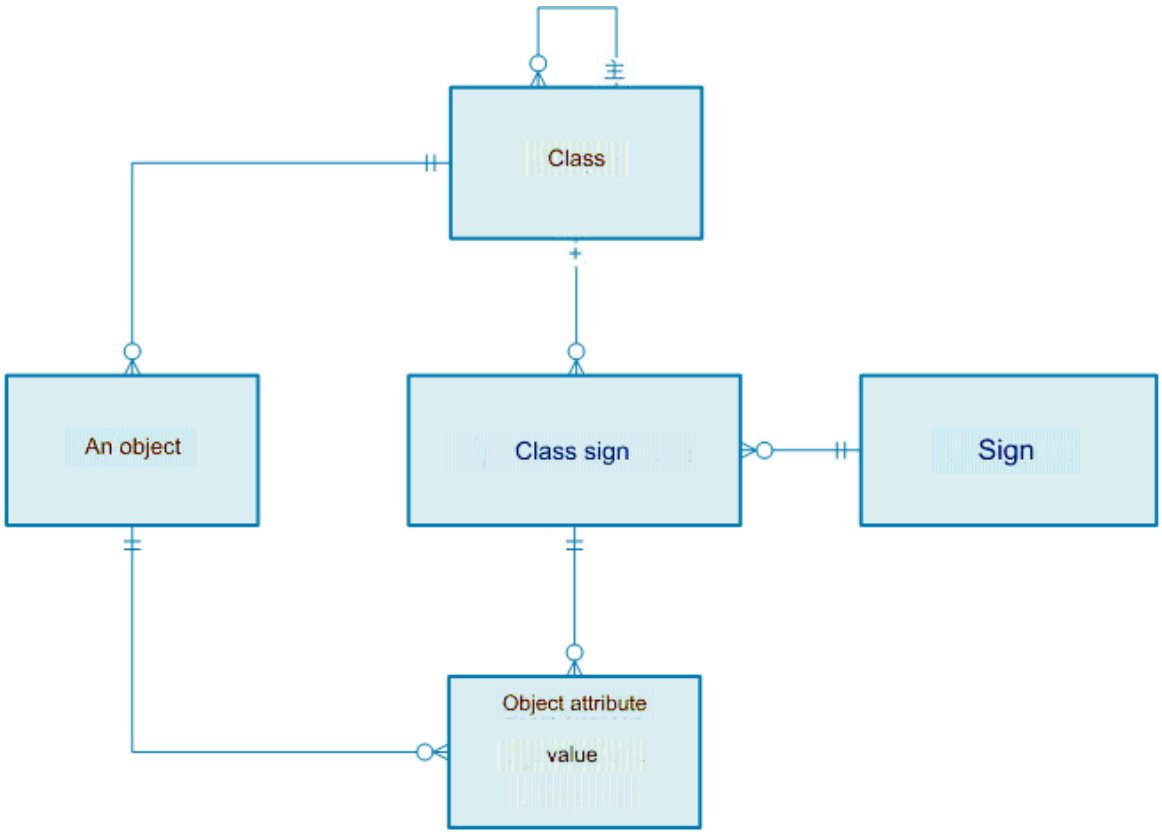

Figure 6. Conceptual data model for using selected sets of feature values for a class

3.3) Using individual analytics

To illustrate this method, a logical data model is more suitable (Figure 7). In this case, it shows that to obtain additional analytics (type of cargo), the analyst may use a classifier of goods already used when preparing transactional transportation documents.

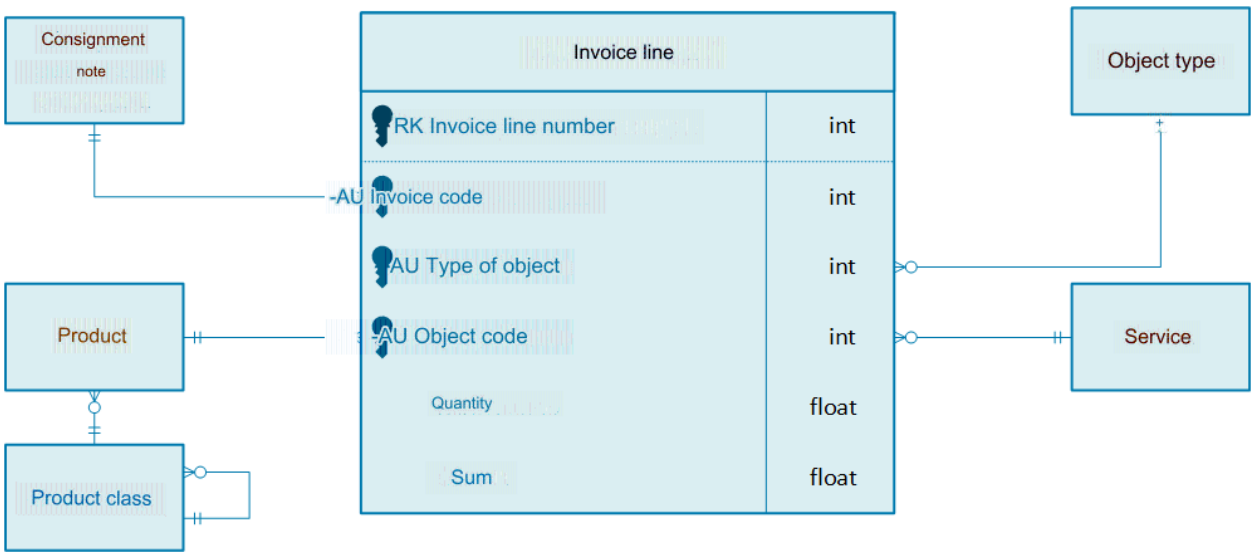

Figure 7. Logical data model for linking services and cargo classes

Conclusion

This paper proposes options for solving problems of classifying works and services, which were successfully tested on projects for developing normative and reference information in a large Russian company with the participation of former MEPhI students [16], [17]IBM Product Master 12.0 platform was used as a tool for classifying and managing reference data in general, which has flexible means of managing the structure of reference data.

Among the classification management tools, the following can be noted:

– Convenient visualization tools, including for multidimensional classifications
– Possibility to import standard or define custom classifications
– View and edit records using different classifications
– Managing mapping of classes from different hierarchies

Additional visualization tools include the following:

– Display of configured workflows for reconciling records by different categories of users in the form of sequence diagrams;
– Display of reference data processing flows when setting up additional data quality check procedures;
– Connecting external visualization tools, such as maps to show the location of a directory object with specified coordinates.

The constructed classifier of works and services has about 400 classes with specified classification features, is used to classify more than 6000 initial records of information systems. 6 subject areas were used as data sources (Purchases, Payments, Costs, Repairs, Investments, Accounting). The directory of works and services is centrally maintained for more than 10 information systems of the holding.

References

1. Kukshev V.I., "Prospects for the Development of Industrial Product Classifiers in the EAEU," 12 12 2023. [Online]. Available: http://www.rgtr.ru/data/events/2023/SMART%2C%2012.12.2023/8.%20%D0%9A%D1%83%D0%BA%D1%88%D0%B5%D0%B2%20%D0%92.%D0%98..pdf.

2. H. Hedden, “Turning a Taxonomy into an Ontology,” 17 11 2021. [Online]. Available: https://www.hedden-information.com/wp-content/uploads/2021/11/Turning-a-Taxonomy-into-an-Ontology.pdf.

3. The Open Group, “TOGAF,” The Open Group, April 2022. [Online]. Available: https://pubs.opengroup.org/togaf-standard/index.html.

4. M.N.Strikhanov, N.N.Degtyarenko, V.V.Pilyugin, E.E. Malikova, N.A. Matveeva, V.D.Adzhtev, A.A.Pasko, “Experience of computer visualization of nanostructures at MEPhI,” Scientific Visualization, vol. 1, no. 1, pp. 1-18, 2009.
5. O.V. Peskova, “On information visualization,” Bulletin of Bauman Moscow State Technical University. Series “Instrument Engineering”, 2012.
6. Ferdio, “Data Viz Project,” Ferdio, [In Internet]. Available: https://datavizproject.com. [Accessed: 30 07 2024].
7. I.K. Romanova, "MODERN METHODS OF MULTIDIMENSIONAL DATA VISUALIZATION: ANALYSIS, CLASSIFICATION, IMPLEMENTATION, APPLICATIONS IN TECHNICAL SYSTEMS," Science and Education: scientific publication of Bauman Moscow State Technical University, No. 3, pp. 133-167, 2016.
8. Ayvazyan S.A., Bezhaeva Z.I., Staroverov O.V., Classification of multivariate observations, Moscow: Statistics, 1974.
9. "Requirements for a hierarchical classification system," [Online]. Available: https://studylib.ru/doc/4675234/trebovaniya-k-ierarhicheskoj-sisteme-klassifikacii. [Accessed: 30 07 2024].
10. K. M. David Mark, Methodology of structural analysis and design: Translated from English, Moscow: Original layout: "Meta-technology", Printing: GMP "First Model Printing House", 1993.
11. Mark Mosley, Michael Brackett, Susan Earley, Deborah Henderson, The DAMA Guide to the Data Management Body of Knowledge, NJ USA: DAMA International, 2010.
12. M. Nizhelskaya, N. Klassen, A. Pavlova, “Entities and Relationships: How and Why System Analysts Create ER Diagrams,” 17 05 2023. [Online]. Available: https://practicum.yandex.ru/blog/chto-takoe-er-diagramma/.
13. "Entity-Relationship Model Notations (ER Diagrams)," 11 09 2021. [Online]. Available: https://pro-prof.com/archives/8126.
14. Dzengelevsky A.E., Mogilat A.S., “Classifiers of the KSSS: now also finished products,” ITime – Information technologies in the fuel and energy complex, v. 2(12), pp. 26-29, 2010.
15. Dzengelevsky A.E., “Levels of Generalization in Accounting for Inventory,” Actual Problems of Humanities and Natural Sciences, Vol. 12 (59), No. I, pp. 82-90, 2013.
16. S. A. Dzengelevsky A. E., “History of the KSSS: Unified Corporate System of Dictionaries and Reference Books,” ITime – Information Technologies in the Fuel and Energy Complex, 2007.
17. H. Sh. Dzengelevsky A.E., “Experience in the creation and development of a corporate system for managing normative and reference information,” Scientific and Methodological Journal “Inter-Industry Information Service”, v. 2(12), pp. 22-26, 2012.