Магжан Жаксылыкович Иген – магистрант Satbayev University, кафедра «Электротехника и космическая техника Телекоммуникации», Алматы, Казахстан; e-mail: 030303500411-m@stud.satbayev.university. ORCID: https://orcid.org/0009-0006-7156-5182.

Еркебулан Амандыкович Тулешов – к. т. н., ассоциированный профессор, кафедра робототехники и технических средств автоматики, Satbayev University, Институт автоматики и информационных технологий, Алматы, Казахстан; e-mail: y.tuleshov@satbayev.university. ORCID: https://orcid.org/0000-0002-9146-7137.

Жанибек Назарбекович Исабеков – PhD, ассоциированный профессор, кафедра робототехники и технических средств автоматики, Satbayev University, институт автоматики и информационных технологий, Алматы, Казахстан; e-mail: z.issabekov@satbayev.university. ORCID: https://orcid.org/0000-0003-2900-8025.

References

Nurbek Pazharbekovich Saparkhodzhayev – PhD, associate professor, Chairman of the management board-Rector, Rudny industrial University, Rudny, Kazakhstan; e-mail: nursp81@gmail.com. ORCID: https://orcid.org/0000-0002-9204-5374.

Anar Khabay* – PhD, associate professor, Satbayev University, Department of Electrical Engineering and Telecommunications of space technology, Almaty, Kazakhstan; e-mail: a.khabay@satbayev.university. ORCID: https://orcid.org/0000-0002-0409-1531.

Magzhan Zhaksylykovich Igen – master's student Satbayev University, Department of Electrical Engineering and Telecommunications of space technology, Almaty, Kazakhstan; e-mail: 030303500411-m@stud.satbayev.university. ORCID: https://orcid.org/0009-0006-7156-5182.

Erkebulan Amandykovich Tuleshov – candidate of Technical Sciences, Associate Professor, Department of robotics and technical means of automation, Satbayev University, Institute of automation and information technologies, Almaty, Kazakhstan; e-mail: y.tuleshov@satbayev.university. ORCID: https://orcid.org/0000-0002-9146-7137.

Zhanibek Nazarbekovich Issabekov – PhD, Associate Professor, Department of robotics and technical means of automation, Satbayev University, Institute of automation and information technologies, Almaty, Kazakhstan; e-mail: z.issabekov@satbayev.university. ORCID: https://orcid.org/0000-0003-2900-8025.

Редакцияға енуі 01.06.2026
Жариялауға қабылданды 17.06.2026

https://doi.org/10.53360/2788-7995-2026-3(23)-5

IRSTI: 28.23.13

M.A. Sambetbayeva[1,3], V.B. Barakhnin[2], A.N. Sultangaziyeva[1,4,*], N.K. Mukazhanov[1,5], R.A. Begim[3]

[1]«Q» University, 050026, Republic of Kazakhstan, Almaty, str. Baizakov 125/185,
[2]Novosibirsk State Technical University,
Russia, Novosibirsk 630087, str. Nemirovicha-Danchenko 136.
[3]L.N. Gumilyov Eurasian National University,
010008, Republic of Kazakhstan, Astana, Satpayev str. 2.
[4]Astana International University,
010000, Republic of Kazakhstan, Astana, Qabanbay Batyr Avenue, 8
[5]Narxoz University, Kazakhstan, Almaty 050035, str. Zhandosov 55.
*e-mail: anar_sultangazieva@aiu.edu.kz

ONTOLOGY-AWARE FOR PHENOTYPE TERM NORMALIZATION AND GROUP ASSIGNMENT

Abstract: *This study addresses the problem of automating the processing of terminological data using ontology-oriented methods of natural language processing. Modern information systems increasingly operate with heterogeneous and poorly structured terms presented in the form of lists or textual descriptions, which complicates their normalization, comparison, and subsequent automatic classification. This problem is particularly relevant in biomedical informatics, where phenotypic descriptions of hereditary diseases require standardized representation for integration into intelligent information systems.*

©M.A. Sambetbayeva, V.B. Barakhnin, A.N. Sultangaziyeva, N.K. Mukazhanov, R.A. Begim, 2026

The paper proposes an ontology-aware NLP algorithm aimed at the normalization of terms and their automatic distribution into a fixed set of semantic groups based on an ontological model. The approach combines methods for entity extraction and normalization, ontological linking, and taxonomic classification, enabling the generation of a formalized and reproducible representation of terminological data. As the semantic basis, the specialized PNGO ontology (Phenotype-based Normalized Group Ontology) is used, implemented in OWL 2 format and defining a fixed set of 23 phenotypic groups anchored to the Human Phenotype Ontology (HPO) hierarchy.

The proposed approach provides a deterministic and interpretable alternative to machine and deep learning methods, ensuring reproducibility and semantic consistency of results without requiring a training stage on annotated data. The effectiveness of the solution is confirmed by experimental evaluation on five hereditary diseases comprising 53 phenotype terms, where the algorithm achieved a macro-average F1-score of 0.919. The resulting structured profile is suitable for integration into information systems, analytical modules, and reproducible computational pipelines.

Key words: *ontology-aware NLP; phenotype normalization; knowledge graphs; semantic grouping; SPARQL; biomedical informatics.*

Introduction

The development of information technologies and the increase in the volume of data is accompanied by an increase in the volume of terminological data used in information, analytical and intellectual systems. Such data is often presented in an unstructured or poorly formalized form, which significantly limits the ability to automatically process, compare and reuse them in computational pipelines. To solve this problem, methods of information search and extraction, natural language processing, machine learning and deep learning algorithms can be used [1-4]. However, the search for information and obtaining specific data related to a particular subject area requires individual research and the search for special approaches. This is because fields such as medicine, law, marketing, etc. have their own terminology, meaning of terms and relationship semantics, as well as specifics of presenting data in textual form. One of the ways to solve the problems of identification and extraction of semantic connections between text data is the ontological approach [5, 6].

One of the urgent tasks of ontology-aware NLP is to develop methods for normalization and semantic binding of terms with controlled dictionaries and ontologies. The use of ontological models allows for the consideration of hierarchical and semantic relationships between the terms, providing a more stable and interpretable classification compared to exclusively statistical approaches.

In this paper, an algorithm for automatic normalization of terms and their distribution on semantic groups based on the ontological model is proposed. Unlike traditional classification methods, the proposed approach uses the taxonomic structure of ontology as the basis for the formation of a reproducible and formalized representation of terminological data.

The goal of the paper is to develop and test the ontology-aware NLP algorithm, which provides automatic normalization of terms and their grouping within the reproducible computational pipelines, focused on the tasks of automation and information technology.

Methods

In modern research on processing phenotypic and biomedical information, machine learning, deep learning and clinical NLP methods are widely used, focused on extracting and normalizing terms from unstructured texts [7-9]. Such approaches, as a rule, use trained neural and transformer models and demonstrate high accuracy rates, but require significant volumes of marked data, stages of learning and post-learning, and are characterized by limited interpretability and difficulties of reproducibility of results [5, 10].

An alternative direction is ontology-aware and knowledge graph-oriented methods in which the ontological structure is used as a direct semantic basis for data processing [3-6]. Within the framework of this work, this approach is used: The proposed ontology-aware NLP method is implemented in the form of a deterministic algorithm and does not involve the use of trained models, providing interpretability, reproducibility and formalized distribution of phenotypic terms on a fixed set of phenotypic groups.

The paper proposes a deterministic method of automatic processing of phenotypic information, based on the use of ontologies and controlled dictionaries. The approach is focused on the normalization of phenotypic terms and their subsequent distribution on a predetermined set of phenotypic groups. The use of ontologies and knowledge graphs as a semantic basis corresponds to modern approaches in the field of knowledge processing and information technology [3, 6, 11].

The study is methodical and algorithmic in nature and does not involve the use of patient data or clinical histories. All operations are performed on open, standardized sources of knowledge, which ensures reproducibility of results and absence of ethical restrictions. As a semantic basis, the specialized applied ontology PNGO (Phenotype-based Normalized Group Ontology), developed for the problems of normalization and grouping of phenotypic features, is used.

The PNGO ontology (Phenotype-based Normalized Group Ontology) was developed specifically for the tasks of normalization and structured grouping of phenotypic features of hereditary diseases. The ontology is implemented in OWL 2 / Turtle format and comprises 9 classes, 8 object properties, and 15 data properties. The classes include: Disease, PhenotypeMention, NormalizedPhenotype, PhenotypeGroup, GroupAssignment, InheritanceStatement, ExtractionRun, SourceRecord, and MatchingMethod. The ontology supports four normalization strategies: dictionary match, fuzzy match, embedding similarity match, and rule-based match.

The ontology defines a fixed set of 23 phenotypic groups anchored to the top-level hierarchy of the Human Phenotype Ontology (HPO). Each group is linked to an anchor HPO identifier that defines its semantic scope. The 23 groups are: Head abnormalities (HP:0000235), Facial abnormalities (HP:0000272), Nervous system abnormalities (HP:0000708), Eye abnormalities (HP:0000479), Ear abnormalities (HP:0000599), Nose abnormalities (HP:0000368), Mouth abnormalities (HP:0000154), Neck abnormalities (HP:0000465), Cardiovascular abnormalities (HP:0001627), Respiratory abnormalities (HP:0002087), Digestive system abnormalities (HP:0025031), Metabolism or homeostasis abnormalities (HP:0001941), Endocrine abnormalities (HP:0000819), Genitourinary abnormalities (HP:0000121), Musculoskeletal abnormalities (HP:0033127), Skeletal abnormalities (HP:0000925), Blood and hematopoietic abnormalities (HP:0001872), Immune system abnormalities (HP:0002716), Integument abnormalities (HP:0001575), Growth abnormalities (HP:0001508), Prenatal development abnormalities (HP:0001199), Neoplasms (HP:0002665), and Mode of inheritance (HP:0000005).

The input data are standardized phenotypic descriptions presented in the form of ontological identifiers, text terms and their synonymous forms, as well as, if any, inheritance type attributes. Ontological sources are considered as immutable reference books and are used exclusively to extract the structure of knowledge and semantic connections. In contrast to the trained models, the ontologically based approach ensures the interpretability and predictability of the results.

PNGO ontology is task-oriented and optimized to distribute phenotypes at a fixed upper level, represented by a set of 23 phenotypic groups reflecting the main clinical-anatomical and functional categories. Ontology is implemented in the form of an OWL model and includes hierarchically organized classes of phenotypic features and properties that ensure their semantic binding with the corresponding groups. The hierarchical structure allows to determine the belonging of phenotypic features to groups based on their position in taxonomy, which ensures the interpretability and reproducibility of the grouping procedure.

Formally, the task is to construct a mapping of a set of phenotypic features

$$P = \{p_1, p_{2,\dots} p_n\} \quad (1)$$

into a set of phenotypic groups

$$G = \{g_1, g_{2,\dots} g_{23}\} \quad (2)$$

as a result, a structured phenotypic profile of the disease is automatically formed. Such a profile represents the distribution of phenotypic features into groups and allows you to move from an unstructured list of terms to a formalized representation suitable for computational processing.

Each phenotypic feature is pre-normalized and associated with the corresponding concept of ontology, which eliminates terminological ambiguity and allows the use of the hierarchy of ontology in analysis. Additionally, a normalized inheritance type attribute can be included in the profile as a separate element, ensuring the integrity of the representation. The quantitative characteristics of the ontology are shown in Figure 1 (see Figure 1).

The figure presents the main quantitative characteristics of the ontology, including the number of classes, axioms, object and annotation properties, as well as hierarchical relationships that reflect the scale and structural complexity of the model used.

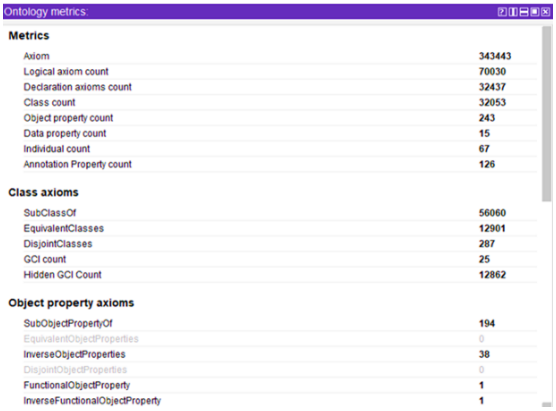


Figure 1 – Protégé Ontology Metrics

Architecture of Methodological Solution

The proposed method is implemented in the form of a sequential ontological-oriented pipeline, which includes the stages of preparation of input data, normalization of terms, ontological binding, distribution of phenotypes into groups and formation of a structured result. The steps are logically independent and can be reproduced separately, which increases the transparency and extensibility of the solution (see Figure 2).

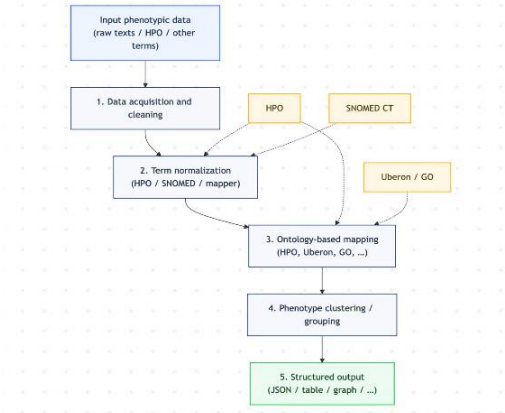


Figure 2 – Ontologically oriented pipeline of normalization and grouping of phenotypic data

Normalization and Ontological Binding

At the normalization stage, input phenotypic terms are brought to the canonical form by bringing the register, removing service symbols, matching synonyms and linking text forms to ontological identifiers. This eliminates terminological variability and ensures unambiguous alignment with the ontological representation.

Next, the ontological binding of phenotypes with classes of PNGO ontology is performed on the basis of coincidence of identifiers, normalized names and synonymous relationships, taking into account hierarchical consistency. In cases of ambiguity, a priority rule based on the level of specificity of the concept in taxonomy is applied.

Distribution by Phenotypic Groups

Distribution of phenotypic features by groups is carried out on the basis of the analysis of the hierarchy of ontology. For each normalized feature, its belonging to one or more phenotypic groups is determined in accordance with predetermined rules. Multiple ownership uses a controlled priority strategy.

Phenotypic groups are represented in PNGO ontology as upper-level classes, which ensures the interpretability and reproducibility of the grouping procedure.

Output View Generation and Reproducibility

The result of the method is a structured phenotypic profile of the disease, presented in machine-readable form (table or JSON), including a list of phenotypic groups, their corresponding signs and a normalized value of the type of inheritance.

The method is completely deterministic and does not contain stochastic components. All processing steps and distribution rules are formalized and can be reproduced using the same input data and ontology versions.

The proposed algorithm is based on the ontologically-based grouping of phenotypic features and the use of the knowledge graph-oriented representation of knowledge [3, 6, 12]. The algorithm is deterministic and is implemented in the form of sequential processing of input phenotypic data using an ontological structure and a fixed set of phenotypic groups.

In the first step, many phenotypic features are extracted from the source of standardized data, presented in the form of ontological identifiers or text terms. If available, information about the type of inheritance of the disease is also extracted (see Formula 1).

Next, the normalization of phenotypic features is performed, including the reduction of terms to canonical form, the elimination of variability and the comparison of text representations with ontological identifiers. As a result, each phenotypic feature is uniquely associated with the corresponding concept of ontology.

After normalization, phenotypes are ontologically linked to the nodes of PNGO ontology, taking into account their position in the hierarchy. In cases of ambiguous correspondence, a priority rule based on semantic specificity and depth of the concept in taxonomy is applied.

For each normalized phenotypic feature, its belonging to phenotypic groups from a fixed set is determined (see Formula 2).

The distribution is based on the analysis of the ontological hierarchy and predefined compliance rules. Multiple potential affiliation uses a fixed ambiguity resolution strategy that provides either unambiguous classification or controlled multi-group affiliation.

Unlike approaches based on learning models and large language models [7,10-13], the proposed algorithm does not require a learning stage and does not contain stochastic components.

The result of the algorithm is a structured phenotypic profile of the disease

$$Profil(D) = \{(G_k, P_k)\} \quad (3)$$

where each phenotypic group is compared a subset of phenotypic features, or an empty set in their absence. The normalized value of the inheritance type is included in the profile as a separate category.

The final view is formed in a machine-readable format (table or hierarchical structure) and can be used for analysis, visualization and integration into intelligent data processing systems. The algorithm provides determinism, interpretability, reproducibility, and extensibility.

Results

As a result of the application of the proposed algorithm for each disease, a structured phenotypic profile is formed, representing the distribution of normalized phenotypic features on a fixed set of phenotypic groups. The algorithm provides a transition from an unstructured list of phenotypes to a formalized and machine-readable representation suitable for subsequent analysis and integration into information systems.

The obtained results demonstrate that the use of the ontological structure allows to correctly take into account the semantic connections between phenotypes and ensure their coordinated distribution across clinical and anatomical groups.

As an illustrative example, disease D is considered, for which a list of phenotypic features in the form of ontological identifiers and text terms is presented in the source of standardized data. After normalization and ontological binding, each phenotypic feature was automatically assigned to the corresponding phenotypic group.

The result of the algorithm is a structured phenotypic profile of the disease, in which phenotypic features are grouped by clinical-anatomical systems. At the same time, for certain groups there may be a lack of phenotypes, which is also an informative result and reflects the specifics of the phenotypic description of a particular disease (see Figure 3).

The figure shows an individual of ontology corresponding to the disease, with automatically associated normalized phenotypic traits and an attribute of the type of inheritance, illustrating the result of the formation of a structured phenotypic profile.

Extraction and representation of the phenotypic profile was performed using requests to the ontological model in the Protégé environment. The distribution of phenotypic features across groups is presented in Table 1. (see Table 1).

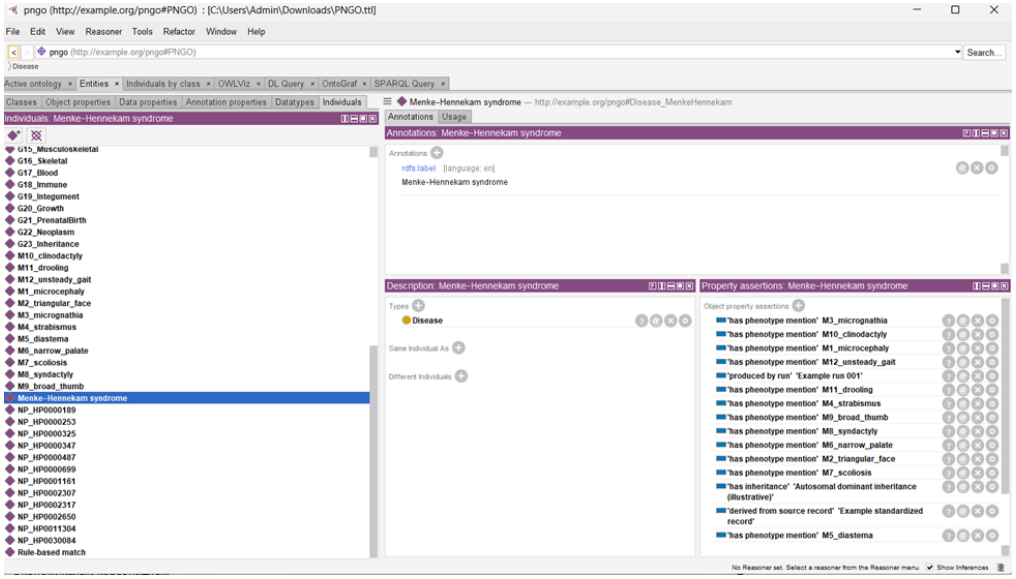


Figure 3 – An example of ontological representation and phenotypic annotation of the disease (Menke-Hennekam syndrome)

Table 1 – An example of a structured phenotypic profile of the disease

№	Phenotypic group	Phenotypic features
1	Abnormalities of head development	Microcephaly
2	Abnormalities of facial development	Hypertelorism
3	Abnormalities of the nervous system	Delayed psychomotor development
4	Abnormalities of the musculoskeletal system	Kyphoscoliosis
5	Abnormalities of the skin	Skin abnormalities
6	Growth abnormalities	Growth retardation
7	Inheritance pattern	Autosomal recessive

Analysis of the results shows that the algorithm correctly distributes phenotypic features to phenotypic groups based on ontological hierarchy. Phenotypes related to different clinical-anatomical systems are not mixed between groups, which confirms the semantic consistency of the method.

In addition, the explicit representation of empty groups allows preserving the integrity of the phenotypic profile and using it for subsequent comparison of diseases, aggregation of data and construction of feature representations.

The graph representation of the formed phenotypic profile of the disease is shown in Figure 4 (see Figure 4).

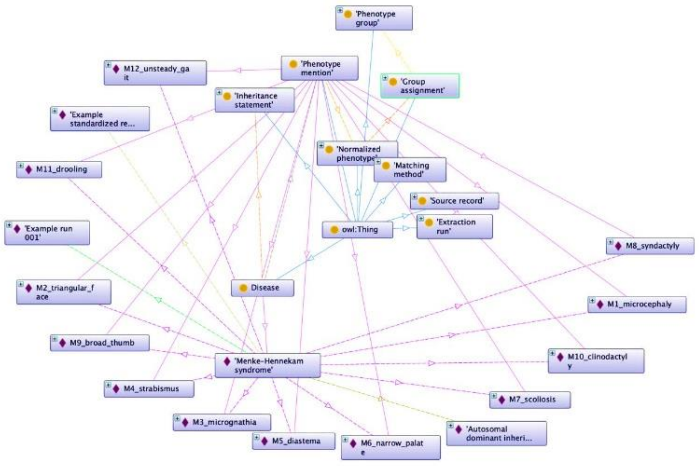


Figure 4 – Graphic representation of the structured phenotypic profile of the disease

The figure shows an ontological graph representation of the disease with nodes corresponding to phenotypic features and type of inheritance, and edges reflecting semantic relationships between them. The graph illustrates the result of the proposed algorithm and demonstrates the formation of a coherent phenotypic profile of the disease.

To confirm the reproducibility and feasibility of the proposed ontological approach, the extraction of the structured phenotypic profile of the disease was carried out using standard SPARQL queries directly to the PNGO ontology.

Queries were performed in the Protégé environment using the Snap SPARQL Query interface and provided a bypass of the ontological graph from the individual disease through associated phenotypic traits normalized in terms of Human Phenotype Ontology (HPO) to the corresponding phenotypic groups.

The formation of the phenotypic profile of the disease is realized not as external data post-processing, but as a direct extraction of knowledge formalized in the ontological model (see Figure 5).

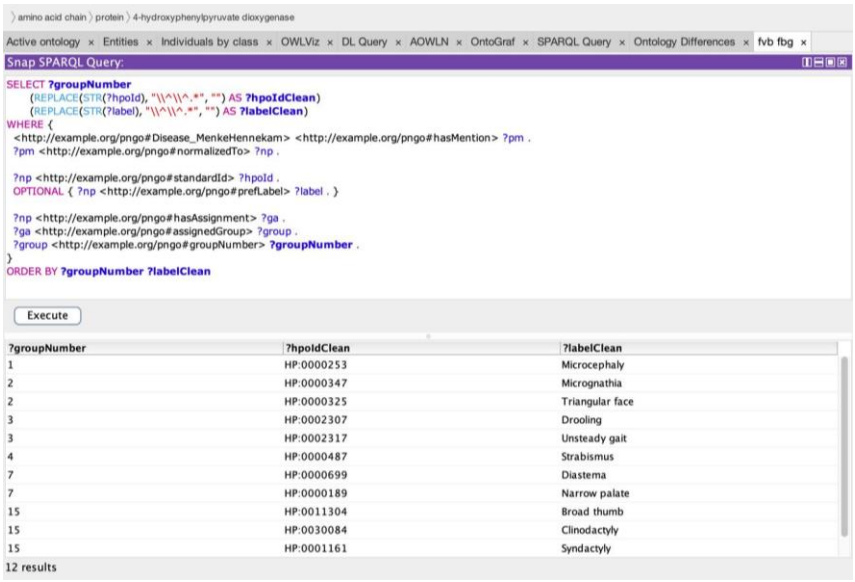


Figure 5 – Result of execution of SPARQL-request for extraction of structured phenotypic profile of Menke-Hennekam syndrome

The obtained results confirm that the proposed algorithm provides automatic formation of a structured phenotypic profile of the disease and can be used in the tasks of intellectual processing of medical information, analysis of hereditary diseases and preparation of data for subsequent computational methods.

To demonstrate the universality of the proposed algorithm, it was applied to four additional hereditary diseases with distinct phenotypic profiles: Marfan Syndrome (OMIM 154700), Down Syndrome (OMIM 190685), Turner Syndrome (OMIM 613577), and Cystic Fibrosis (OMIM 219700). For each disease, phenotypic features were extracted from standardized data sources in HPO identifier format, normalized, and automatically assigned to the corresponding phenotypic groups of the PNGO ontology. The resulting structured profiles are summarized in Table 2.

To quantitatively evaluate the performance of the proposed algorithm, a pilot experiment was conducted on five hereditary diseases comprising 53 phenotype terms in total. For each term, the algorithm output was compared against a gold-standard group label assigned according to the HPO top-level organ system hierarchy. The evaluation metrics are presented in Table 3.

The results across all five diseases confirm that the algorithm correctly assigns phenotypic features to semantically appropriate groups based on the ontological hierarchy, achieving a macro-average F1-score of 0.919. The lower performance observed for Turner Syndrome (F1 = 0.778) is attributable to phenotypic overlap between the Endocrine and Genitourinary groups, reflecting a known ambiguity in HPO taxonomy. These results demonstrate the robustness and generalizability of the proposed approach across diseases with diverse phenotypic profiles.

Table 2 – Structured phenotypic profiles of additional diseases

Disease (OMIM)	Phenotypic group	Phenotypic features (HPO)
Marfan Syndrome (154700)	Cardiovascular	Mitral valve prolapse (HP:0001634); Aortic root aneurysm (HP:0002616)
	Eye	Ectopia lentis (HP:0001083); Deeply set eye (HP:0000490)
	Skeletal	Scoliosis (HP:0002650); Pectus carinatum (HP:0000768)
Down Syndrome (190685)	Nervous system	Seizures (HP:0001250); Globaldevelopmental delay (HP:0001263); Delayed speech (HP:0000750)
	Cardiovascular	Atrial septal defect (HP:0001631); Ventricular septal defect (HP:0001629)
	Growth	Failure to thrive (HP:0001508)
Turner Syndrome (613577)	Growth	Short stature (HP:0000876)
	Cardiovascular	Bicuspid aortic valve (HP:0001631)
	Integument	Hypohidrosis (HP:0000966); Hyperpigmentation (HP:0000953)
Cystic Fibrosis (219700)	Respiratory	Bronchiectasis (HP:0002110); Asthma (HP:0002099); Emphysema (HP:0002097)
	Digestive	Exocrine pancreatic insufficiency (HP:0001738); Malabsorption (HP:0002024)
	Endocrine	Diabetes mellitus (HP:0000819)

Table 3 – Quantitative evaluation of the PNGO algorithm (Precision, Recall, F1-score)

Disease	Terms	TP	FP	Prec.	Recall	F1
Menke–Hennekam syndrome	12	12	0	1.000	1.000	1.000
Marfan Syndrome	10	9	1	0.900	0.900	0.900
Down Syndrome	12	11	1	0.917	0.917	0.917
Turner Syndrome	9	7	2	0.778	0.778	0.778
Cystic Fibrosis	10	10	0	1.000	1.000	1.000
Overall (Macro-average)	53	49	4	0.919	0.919	0.919

Discussion

In this paper, a deterministic ontologically based algorithm for the formation of a structured phenotypic profile is proposed, based on the display of phenotypic features in a fixed set of upper-level phenotypic groups. The obtained results are correlated with modern research in the field of ontology-aware NLP, ontological modeling and knowledge graph-oriented approaches.

A structured comparison of the proposed PNGO algorithm with existing approaches (ML/DL, LLM-based, knowledge graph-based, and ontology-aware NLP) is presented in Table 4.

Table 4 – Comparison of the proposed PNGO algorithm with existing approaches

Criterion	ML/DL	LLM-based	KG-based	Proposed (PNGO)	Ontology NLP
Requires training data	Yes (large)	Yes (large)	No	No	No
Interpretability	Low	Low–Medium	Medium	High	Medium
Reproducibility	Stochastic	Stochastic	Deterministic	Deterministic	Deterministic
Semantic grouping	Implicit	Implicit	Partial	Explicit (23 gr.)	Partial
Ontology as core basis	No	Partial	Yes	Yes	Yes
Structured profile output	No	No	Partial	Yes	No
Normalization to HPO	Partial	Partial	Yes	Yes	Yes
F1-score (this study)	–	–	–	0.919	–

Unlike the methods of machine and deep learning, widely used to extract phenotypic information from texts [7, 8, 9]. the proposed algorithm does not require a learning stage and provides interpretability, reproducibility and predictability of the results. The ontological structure is used not as an infrastructure layer, but as a direct semantic basis of the grouping algorithm, which allows to form an interpreted profile without the involvement of statistical models [3, 4, 6].

An additional level of abstraction is proposed in the form of a fixed set of phenotypic groups, which allows to move from flat lists of phenotypes to a structured representation and increase the comparability of profiles. Semantic aggregation is formalized as an explicit mapping of $F: P \rightarrow G$, which enhances the interpretability and reproducibility of the method.

The novelty of this work lies in:

1) deterministic ontology-based grouping without ML,

2) introduction of PNGO ontology with fixed 23 groups,

3) formal mapping F: P → G for phenotype structuring.

The limitations of the approach include the dependence of results on the quality of the source data and the ontology used, as well as the fixed set of groups, which requires possible adaptation for other subject areas. Promising areas of further research are the integration of the algorithm with methods of automatic extraction of terms from unstructured texts [7,8] and the development of hybrid solutions that combine deterministic ontological grouping with the capabilities of modern language models [5, 10].

Conclusion

The article presents ontology-aware NLP-algorithm of automatic normalization of terms and their distribution on a fixed set of semantic groups based on the ontological model. The proposed approach provides a transition from unstructured terminological data to a formalized and machine-readable representation suitable for use in automated information systems.

Experimental testing showed that the use of the ontological structure allows to provide semantically consistent classification of terms, as well as to preserve the reproducibility of results due to the explicit formalization of distribution rules in ontology. The resulting structured profile can be used for subsequent analysis, data aggregation and integration into knowledge-based systems.

The proposed method is focused on the problems of automation and information technology and can be extended to handle various subject areas using controlled dictionaries, taxonomies and ontologies. Further development of the work is associated with the scaling of the algorithm and its integration into complex reproducible NLP-computational pipelines.

References

1. Uyghur–Kazakh–Kirghiz text keyword extraction based on morpheme segmentation / S. Parhat et al // Information. 2023. – Vol. 14, № 5. – P. 283. https://doi.org/10.3390/info14050283.
2. Hoehndorf R. The role of ontologies in biological and biomedical research: A functional perspective / R. Hoehndorf, P.N. Schofield, G.V. Gkoutos // Briefings in Bioinformatics. – 2015. – Vol. 16, № 6. – P. 1069-1080. https://doi.org/10.1093/bib/bbv011.
3. A review on knowledge graphs for healthcare: Resources, applications, and promises / H. Cui et al // Journal of Biomedical Informatics. – 2025. – Vol. 169. – P. 104861. https://doi.org/10.1016/j.jbi.2025.104861.
4. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach / J.-B. Lamy et al // Artificial Intelligence in Medicine. – 2019. – Vol. 94. – P. 42-53. https://doi.org/10.1016/j.artmed.2019.01.001.
5. Opportunities and challenges for ChatGPT and large language models in biomedicine and health / S. Tian et al // Briefings in Bioinformatics. – 2024. – Vol. 25, № 1. – P. bbad493. https://doi.org/10.1093/bib/bbad493.
6. Riquelme-García A. Annotation of biological samples data to standard ontologies with support from large language models / A. Riquelme-García, J. Mulero-Hernández, J.T. Fernández-Breis // Computational and Structural Biotechnology Journal. – 2025. – Vol. 27. – P. 2155-2167. https://doi.org/10.1016/j.csbj.2025.02.018.
7. Deep learning in clinical natural language processing: A methodical review / S. Wu et al // Journal of the American Medical Informatics Association. – 2020. – Vol. 27, № 3. – P. 457-470. https://doi.org/10.1093/jamia/ocz200.
8. Arora A. Named entity recognition for clinical text: A review / A. Arora, E. Nyberg, C.P. Rosé // Artificial Intelligence in Medicine. – 2021. – Vol. 113. – P. 102039. https://doi.org/10.1016/j.artmed.2021.102039.
9. Clinical natural language processing in languages other than English: Opportunities and challenges / A. Névéol et al // Journal of Biomedical Semantics. – 2020. – Vol. 11. – P. 12. https://doi.org/10.1186/s13326-020-00239-7.

10. Lu Z. Large language models in biomedicine and health / Z. Lu, Y. Peng // Journal of the American Medical Informatics Association. – 2024. – Vol. 31, № 8. – P. 1801-1811. https://doi.org/10.1093/jamia/ocae089.
11. BioPortal: An open community resource for sharing, searching, and utilizing biomedical ontologies / J. Vendetti et al // Bioinformatics. – 2025. – Vol. 41, № 2. – P. gkaf402. https://doi.org/10.1093/bioinformatics/gkaf4020.
12. Biomedical knowledge graph–optimized prompt generation for large language models / K. Soman et al // Bioinformatics. – 2024. – Vol. 40, № 9. – P. btae560. https://doi.org/10.1093/bioinformatics/btae560.
13. A survey on unifying large language models and knowledge graphs in biomedicine / R. Xu et al // Bioinformatics. – 2025. – Vol. 41, No. 1. – P. btaf018. – DOI: 10.1093/bioinformatics/btaf018
14. Nicholson D. N., Greene C. S. Constructing knowledge graphs and their biomedical applications // Computational and Structural Biotechnology Journal. – 2020. – Vol. 18. – P. 1414-1428. – DOI: 10.1016/j.csbj.2020.05.017
15. Hogan A., Blomqvist E., Cochez M. [et al.]. Knowledge graphs // ACM Computing Surveys. – 2021. – Vol. 54, No. 4. – P. 1-37. – DOI: 10.1145/3447772

Funding: *This research has been funded by the Committee of Science of the Ministry of Science and higher Education of the Republic of Kazakhstan Grant AP26105113.*

М.А. Самбетбаева[1,3], В.Б. Барахнин[2], А.Н. Сұлтанғазиева[1,4*], Н.К. Мукажанов[1,5], Р.А. Бегім[3]
[1]«Q» University,
050026, Қазақстан Республикасы, Алматы қ., Байзақов көш., 125/185
[2]Новосибирск мемлекеттік техникалық университеті,
Ресей, Новосибирск қ., 630087, Немирович-Данченко көш., 136
[3]Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
010008, Қазақстан Республикасы, Астана қ., Сәтбаев көш., 2
[4]Astana International University,
010000, Қазақстан Республикасы, Астана қ., Қабанбай батыр даңғ., 8
[5]Narxoz University,
050035, Қазақстан Республикасы, Алматы қ., Жандосов көш., 55
*e-mail: anar_sultangazieva@aiu.edu.kz

ОНТОЛОГИЯҒА БАҒЫТТАЛҒАН ФЕНОТИП ТЕРМИНДЕРІН НОРМАЛДАНДЫРУ МЕН ТОПТАСТЫРУҒА АРНАЛҒАН ӘДІСТЕР

Бұл жұмыста терминологиялық деректерді өңдеуді автоматтандыру мәселесі онтологияға бағытталған табиғи тілдерді өңдеу әдістерін қолдану арқылы қарастырылады. Қазіргі ақпараттық жүйелер әртүрлі және әлсіз құрылымдалған терминдермен жұмыс істейді, олар көбінесе тізімдер немесе мәтіндік сипаттамалар түрінде ұсынылады, бұл олардың нормалдануын, салыстырылуын және кейінгі автоматты жіктелуін қиындатады. Бұл мәселе биомедициналық информатикада ерекше өзектілікке ие, өйткені тұқым қуалайтын аурулардың фенотиптік сипаттамалары интеллектуалды ақпараттық жүйелерге интеграция үшін стандартталған көріністі талап етеді.

Мақалада онтологияға негізделген NLP алгоритмі ұсынылған, ол терминдерді нормалдандыруға және оларды онтологиялық модель негізінде алдын ала анықталған 23 семантикалық топ жиынтығына автоматты түрде бөлуге бағытталған. Ұсынылған тәсіл мәндерді шығару және нормалдандыру, онтологиялық байланыстыру және таксономиялық жіктеу әдістерін біріктіре отырып, терминологиялық деректердің формалданған және қайта жаңғыртылатын көрінісін қамтамасыз етеді. Семантикалық негіз ретінде OWL 2 форматында жүзеге асырылған және Human Phenotype Ontology (HPO) иерархиясына байланған 23 фенотиптік топтың бекітілген жиынтығын анықтайтын PNGO (Phenotype-based Normalized Group Ontology) мамандандырылған онтологиясы қолданылады.

Ұсынылған тәсіл машиналық және терең оқыту әдістеріне детерминирленген және интерпретацияланатын балама болып табылады, таңбаланған деректерде оқыту кезеңінсіз нәтижелердің қайта жаңғыртылуын және семантикалық сәйкестігін қамтамасыз етеді. Шешімнің тиімділігі бес тұқым қуалайтын ауру бойынша (53 фенотиптік термин) эксперименттік бағалаумен расталған, мұнда алгоритм макро-орташаланған F1 = 0,919 көрсеткішіне қол жеткізді. Нәтижелі құрылымдалған профиль ақпараттық жүйелерге, аналитикалық модульдерге және қайта орындалатын есептеу конвейерлеріне енгізуге жарамды.

Түйін сөздер: *онтологияға негізделген NLP; фенотиптерді нормалдандыру; білім графтары; семантикалық топтастыру; SPARQL; биомедициналық информатика.*

М.А. Самбетбаева[1,3] , В.Б. Барахнин[2] , А.Н. Сұлтанғазиева[1,4*], Н.К. Мукажанов[1,5], Р.А. Бегім[3]
[1]«Q» University, 050026, Қазақстан Республикасы, Алматы қ., Байзақов көш., 125/185
[2]Новосибирск мемлекеттік техникалық университеті,
Ресей, Новосибирск қ., 630087, Немирович-Данченко көш., 136
[3]Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
010008, Қазақстан Республикасы, Астана қ., Сәтбаев көш., 2
[4]Astana International University,
010000, Қазақстан Республикасы, Астана қ., Қабанбай батыр даңғ., 8
[5]Narxoz University,
050035, Қазақстан Республикасы, Алматы қ., Жандосов көш., 55
*e-mail: anar_sultangazieva@aiu.edu.kz

ОНТОЛОГИЧЕСКИ ОРИЕНТИРОВАННАЯ НОРМАЛИЗАЦИЯ ТЕРМИНОВ ФЕНОТИПОВ И ПРИСВОЕНИЕ ГРУПП

В данной работе рассматривается задача автоматизации обработки терминологических данных с использованием онтологически-ориентированных методов обработки естественного языка. Современные информационные системы всё чаще работают с разнородными и слабо структурированными терминами, представленными в виде списков или текстовых описаний, что существенно усложняет их нормализацию, сопоставление и последующую автоматическую классификацию. Особую актуальность данная проблема приобретает в биомедицинской информатике, где фенотипические описания наследственных заболеваний требуют стандартизированного представления для интеграции в интеллектуальные информационные системы.

В статье предложен ontology-aware NLP алгоритм, направленный на нормализацию терминов и их автоматическое распределение по фиксированному набору семантических групп на основе онтологической модели. Подход объединяет методы извлечения и нормализации сущностей, онтологического связывания и таксономической классификации, обеспечивая формализованное и воспроизводимое представление терминологических данных. В качестве семантической основы используется специализированная онтология PNGO (Phenotype-based Normalized Group Ontology), реализованная в формате OWL 2 и определяющая фиксированный набор из 23 фенотипических групп, привязанных к иерархии Human Phenotype Ontology (HPO).

Предложенный подход представляет собой детерминированную и интерпретируемую альтернативу методам машинного и глубокого обучения, обеспечивая воспроизводимость и семантическую согласованность результатов без необходимости этапа обучения на размеченных данных. Эффективность решения подтверждена экспериментальной оценкой на пяти наследственных заболеваниях (53 фенотипических термина), где алгоритм достиг макро-усреднённого показателя F1 = 0,919. Результирующий структурированный профиль пригоден для интеграции в информационные системы, аналитические модули и воспроизводимые вычислительные конвейеры.

Ключевые слова: *онтологически-ориентированный NLP; нормализация фенотипов; графы знаний; семантическая группировка; SPARQL; биомедицинская информатика.*

Сведения об авторах

Мадина Аралбаевна Самбетбаева – Phd, ассоциированный профессор кафедры информационных систем, Евразийский национальный университет имени Л.Н. Гумилева, ведущий научный сотрудник «Q» University, Республика Казахстан Астана; e-mail:sambetbayeva_ma_1@enu.kz. ORCID:https://orcid.org/0000-0001-9358-1614.

Владимир Борисович Барахнин – доктор технических наук; заведующий лабораторией информационных ресурсов Федерального исследовательского центра информационных и вычислительных технологий, профессор факультета прикладной математики и информатики Новосибирского государственного технического университета, Новосибирск, Россия; e-mail: bar@ict.nsc.ru. ORCID: https://orcid.org/0000-0003-3299-0507.

Анар Несіпбекқызы Сұлтанғазиева* – магистр информатики, докторант Высшая школа информационных технологий и инженерии, Международный университет Астана, младший сотрудник «Q» University, Республика Казахстан Астана; e-mail:anar_sultangazieva@aiu.edu.kz. ORCID: https://orcid.org/0009-0009-9038-5234.

Нуржан Какенович Мукажанов – руководитель проекта «Q» University, PhD, ассоциированный профессор Школы цифровых технологий, Университет Нархоз Республика Казахстан Алматы; e-mail:nurzhan.mukazhanov@narxoz.kz. ORCID: https://orcid.org/0000-0003-4835-5751.

Раушан Аянқызы Бегім – магистрант кафедры информационных систем, Евразийский национальный университет имени Л.Н. Гумилева, Республика Казахстан Астана; e-mail:rsn.alst@gmail.com. ORCID: https://orcid.org/0009-0008-9859-5300.

Авторлар туралы мәліметтер

Мадина Аралбайқызы Самбетбаева – PhD, Ақпараттық жүйелер кафедрасының қауымдастырылған профессоры, Л. Н. Гумилев атындағы Еуразия ұлттық университеті, "Q" University жетекші ғылыми қызметкері, Қазақстан Республикасы Астана; e-mail:sambetbayeva_ma_1@enu.kz. ORCID: https://orcid.org/0000-0001-9358-1614.

Владимир Борисович Барахнин – техника ғылымдарының докторы; Ақпараттық және есептеу технологиялары жөніндегі федералдық зерттеу орталығының Ақпараттық ресурстар зертханасының меңгерушісі; Новосібір мемлекеттік техникалық университетінің Қолданбалы математика және информатика факультетінің профессоры, Новосібір, Ресей; e-mail:bar@ict.nsc.ru. ORCID: https://orcid.org/0000-0003-3299-0507.

Анар Несіпбекқызы Сұлтанғазиева* – информатика магистрі, Ақпараттық технологиялар және инженерия жоғары мектебінің докторанты, Астана халықаралық университеті, «Q» University кіші ғылыми қызметкері, Қазақстан Республикасы Астана; e-mail:anar_sultangazieva@aiu.edu.kz. ORCID: https://orcid.org/0009-0009-9038-5234.

Нұржан Кәкенұлы Мукажанов – «Q» University жоба жетекшісі, PhD, Цифрлық технологиялар мектебінің қауымдастырылған профессоры, Нархоз университеті Алматы; e-mail: nurzhan.mukazhanov@narxoz.kz. ORCID: https://orcid.org/0000-0003-4835-5751.

Раушан Аянқызы Бегім – Ақпараттық жүйелер кафедрасының магистранты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Қазақстан Республикасы Астана; e-mail: rsn.alst@gmail.com. ORCID:https://orcid.org/0009-0008-9859-5300.

Information about the authors

Madina Aralbaevna Sambetbaeva – PhD, associate professor of the department of Information systems, L.N. Gumilyov Eurasian National University, leading researcher at Q University, Republic of Kazakhstan Astana; e-mail: sambetbayeva_ma_1@enu.kz. ORCID: https://orcid.org/0000-0001-9358-1614.

Vladimir Borisovich Barakhnin – Doctor of technical sciences; head of the laboratory of Information resources at the federal research center for information and computational technologies; Professor of the faculty of Applied mathematics and informatics at Novosibirsk State Technical University, Novosibirsk, Russia; e-mail:bar@ict.nsc.ru. ORCID: https://orcid.org/0000-0003-3299-0507.

Anar Nesipbekkyzy Sultangaziyeva* – master of computer science, doctoral student at the Higher School of Information Technologies and Engineering, Astana International University, junior researcher at “Q” University, Republic of Kazakhstan Astana; e-mail: anar_sultangazieva@aiu.edu.kz. ORCID: https://orcid.org/0009-0009-9038-5234.

Nurzhan Kakenovich Mukazhanov – project leader at “Q” University, PhD, associate professor of the School of Digital Technologies, Narxoz University Republic of Kazakhstan Almaty; e-mail: nurzhan.mukazhanov@narxoz.kz. ORCID: https://orcid.org/0000-0003-4835-5751.

Raushan Ayankyzy Begim – master’s student of the Department of Information systems, L.N. Gumilyov Eurasian National University, Republic of Kazakhstan Astana; e-mail: rsn.alst@gmail.com. ORCID: https://orcid.org/0009-0008-9859-5300.

Received 30.03.2026
Revised 26.05.2026
Accepted 28.05.2026