DOI: 10.15514/ISPRAS-2026-38(4)-8

Automated Event Logs Generation Based on Unstructured Internet Sources for Process Analysis Tasks

K. D. Voronova, ORCID: 0009-0006-4972-5485 <KDVoronova@edu.hse.ru>
L. N. Lyadova, ORCID: 0000-0001-5643-747X <LLyadova@hse.ru>

HSE University, Perm Branch,
38, Studencheskaya St., Perm, 614070, Russia.

Abstract. This paper presents an approach to automated structuring event-related information extracted from unstructured textual Internet sources for process mining tasks. In many practical cases, information on events associated with various processes is not presented in the form of ready-made event logs, but is distributed among news publications, reports and other textual materials. This significantly complicates applying process analysis tools that use data presented in certain formats (for example, XES), since it requires searching for relevant texts, interpreting and transforming them into a structured view suitable for analysis with using existing Process Mining tools. The task is even more complicated if it is necessary to extract additional attributes associated with described events, which makes it possible to identify deeper regularities and patterns that characterize the processes under study, to use new methods for analyzing processes that take these characteristics into account. The proposed approach is based on combination of ontological representations of domain knowledge and information on data sources and large language models, generative artificial intelligence. The ontological layer is used to describe domains of the studied processes, open sources of information about them, user queries and results of processing the found information, while generative models provide searching for relevant information, extraction of data on events and their primary processing and structuring. To implement the approach, a general solution architecture has been proposed that allows automating the receipt of text data from various sources based on user requests generated using ontology, preprocessing the received data using generative models, structuring and normalizing the extracted event data, as well as preparing data for generating event logs in specifying formats. To illustrate possibilities of the proposed approach, a controlled experimental scenario is considered. Experiments have shown that the described software solution allows researchers to find relevant text materials, highlight events and their attributes, as well as form a structured view suitable for further logging events, transformation data into event log. At the same time, the results showed necessity additional normalization of data and expert interpretation of part of the extracted information.

Keywords: information retrieval; natural language processing; text mining; event information extraction; event logging; ontology-driven approach; AI-based approach; generative models; large language model; process mining.

For citation: Voronova K.D., Lyadova L.N. Automated Generation of Event Logs for Process Analysis Tasks Based on Unstructured Internet Sources. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 4, part 1, 2026, pp. 153-170. DOI: 10.15514/ISPRAS-2026-38(4)-8.

Автоматизированное формирование журналов событий на основе неструктурированных Интернет-источников для задач анализа процессов

К.Д. Воронова, ORCID: 0009-0006-4972-5485 <KDVoronova@edu.hse.ru>
Л.Н. Лядова, ORCID: 0000-0001-5643-747X <LLyadova@hse.ru>

Национальный исследовательский университет «Высшая школа экономики»,
НИУ ВШЭ – Пермь,
Россия, 614070, г. Пермь, ул. Студенческая, 38.

Аннотация. В статье представлен подход к автоматизированному структурированию информации о событиях, извлекаемой из неструктурированных текстовых Интернет-источников, для задач углублённого анализа процессов. Во многих практических случаях сведения о событиях, происходящих в рамках различных процессов, представлены не в виде готовых журналов событий, а распределены по новостным публикациям, отчётам и другим текстовым материалам. Это существенно усложняет применение средств анализа процессов, использующих данные, представленные в определённых форматах (например XES), поскольку требует выполнить поиск релевантных текстов, их интерпретацию и преобразование в структурированное представление, пригодное для анализа с помощью существующих средств Process Mining. Задача ещё более усложняется, если необходимо извлекать дополнительные атрибуты, связанные с описанными событиями, что даёт возможность выявить более глубокие закономерности, характеризующие исследуемые процессы, применять новые методы анализа процессов, учитывающие эти характеристики. Предлагаемый подход основан на сочетании онтологического представления знаний о предметных областях и источниках данных и больших языковых моделей, генеративного искусственного интеллекта. Онтологический слой используется для описания предметных областей исследуемых процессов, открытых источников информации о них, пользовательских запросов и результатов обработки найденной информации, тогда как генеративные модели обеспечивают поиск релевантной информации, извлечение данных о событиях и их первичную обработку и структурирование. Для реализации подхода предложена общая архитектура решения, позволяющего автоматизировать получение текстовых данных из различных источников на основе пользовательских запросов, сформированных с использованием онтологии, предобработку полученных данных с использованием генеративных моделей, структурирование и нормализацию извлечённых данных о событиях, а также подготовку данных для формирования журналов событий в заданных форматах. Для иллюстрации возможностей предложенного подхода рассматривается контролируемый экспериментальный сценарий. Проведённые эксперименты показали, что описанное программное решение позволяет находить релевантные текстовые материалы, выделять события и их атрибуты, а также формировать структурированное представление, пригодное для последующего преобразования в журнал событий. Вместе с тем результаты показали необходимость дополнительной нормализации данных и экспертной интерпретации части извлечённой информации.

Ключевые слова: поиск информации; обработка текстов на естественных языках; интеллектуальный анализ текстов; извлечение информации о событиях; протоколирование событий; подход, основанный на онтологии; подход на основе ИИ; генеративные модели; большая языковая модель; контроль и анализ процессов.

Для цитирования: Воронова К.Д., Лядова Л.Н. Автоматизированное формирование журналов событий для задач анализа процессов на основе неструктурированных Интернет-источников. Труды ИСП РАН, том 38, вып. 4, часть 1, 2026 г., стр. 153–170 (на английском языке). DOI: 10.15514/ISPRAS-2026-38(4)-8.

1. Introduction

In recent years, process analysis tasks have been addressed in an increasingly broad range of domains. Process Mining methods are applied not only to traditional domains with event logs received from enterprise information systems (operating systems, database management systems,

etc.), but also to processes whose data is presented in textual documents, open databases, news publications, analytical materials, and other digital sources [1-3].

Another feature of these research areas is that additional attributes specific to the appropriate domains are found for events. Analyzing the values of these additional event attributes allows researchers investigating the dynamics of the behavior of complex systems and evaluating changes of the system states defining with these additional attributes. The papers [4-6] consider the issues of analyzing processes with methods taking into account the values of additional attributes but not only mandatory event attributes like case identifier, activity type, and timestamp.

An approach providing generation of events logs according to the user-defined rules based on the analysis of the values of additional attributes, their dynamics is suggested in the papers [7-8]. Event records that reflect these dynamics are included in the formed event log. To describe the rules, a domain specific language has been developed. Software tools have been implemented that provide the ability to create rules with no-code user interface and to present user's rules, describing events, in the system ontology and to interpret them when analyzing attribute values and generating event logs.

In many practical situations, however, the information required for analysis is not available in the form of ready-made event logs but is contained in unstructured textual sources. Under such conditions, process analysis requires information retrieval, interpretation, and transformation into a structured representation [9]. This problem is especially important in the analysis of global processes, where event-related information is distributed across heterogeneous open sources [10-12].

Studies on event analysis based on Internet sources and Process Mining methods show the promise of this direction, while also indicating the need for specialized means of searching, representing, and interpreting event information [10-11]. Ontology-driven approaches make it possible to formalize domain structure, ensure consistent event descriptions, and support knowledge reuse [12-13]. At the same time, generative models open new possibilities for automating information retrieval, extraction, and structuring, although they still require control over the correctness of the obtained results [14-15].

Therefore, there is a clear need to design an intelligent software solution that automates the full cycle of working with event information from unstructured textual sources, from text retrieval to event logging in a form suitable for using process mining tools to solve tasks of process control and analysis.

Presented research project aims at developing an approach to the automated structuring of event information from unstructured textual sources for process analysis tasks and implementing this approach as a research prototype.

The study addresses the following tasks:

1) analysis of existing approaches to finding, extracting and structuring event information from textual sources;
2) developing the general architecture of an ontology-driven and AI-based prototype;
3) creating the ontological and intelligent basis of the proposed solution;
4) experimenting for demonstration of the capabilities of the proposed approach and for evaluation of the results.

2. Text Mining and Event Information Structuring: Related Work

Research on processing textual data for events and process analysis covers event information retrieval, extraction, structuring, and preservation of results for reuse. These tasks are addressed by classical NLP methods (NLP – Natural Language Processing), knowledge-driven approaches based on semantic models and ontologies, and AI-based solutions (AI – Artificial Intelligence) using generative language models (LLM – Large Language Models) [9, 14-17].

Traditional approaches rely on rules, templates, semantic annotation, and NLP methods. They are interpretable and effective in well-formalized domains, but when applied to heterogeneous sources, variable linguistic constructions, and implicitly expressed events, they often require substantial tuning and provide only a limited degree of automation [9]. Studies on event analysis based on Internet sources also show that the required data are usually not available as ready-made event logs and must first be retrieved, interpreted, and transformed [10-11].

Knowledge-driven approaches, especially ontology-based ones, make it possible to formalize domain structure, define key entities and relations, support search extension through semantic links, and preserve processing results for further reuse [12-13, 18-20]. AI-based approaches, particularly those using generative language models, provide greater flexibility and a higher degree of automation in searching, extracting, and structuring information from texts, although their output depends strongly on query formulation, context completeness, and model characteristics and therefore still requires interpretation and control [14-15, 21-22].

Thus, knowledge-driven and AI-based approaches address various aspects of the same problem and can complement each other. Ontologies support domain modeling, search, interpretation, and preservation of results, whereas generative models support retrieval, extraction, and initial structuring of event information from unstructured textual sources [12]-[15], [20]–[22]. The main characteristics of these approaches are summarized in Table 1.

Table 1. Comparison of approaches to searching, extracting, and structuring event information.

Approach	Strengths	Limitations	Role in solving the problem
Rules and templates	Predictable in narrow formalized tasks	Require manual tuning; weak for variable texts	Suitable for repetitive and well-structured cases
NLP methods	Support extraction of entities, events, and relations	Depending on preprocessing quality and domain adaptation	Used for basic extraction and semantic annotation
Ontology-driven approaches	Formalize domain structure, support consistency and knowledge reuse	Require ontology design and maintenance	Support search, interpretation, and structuring
Generative models	Flexible for unstructured texts, provide high automation	Require verification, sensitive to query and context	Used for retrieval, extraction, and initial structuring
Hybrid approaches	Combine semantic consistency with flexibility and automation	More complex to design and integrate	Most promising for comprehensive problem solving

Therefore, the literature review suggests that the most promising direction for processing event information from unstructured texts is the combined use of domain knowledge and modern language models. An approach to software development based on ontology-driven implementation integrated with generative AI is preferable.

3. Proposed Approach: Automated Event Logs Generation Based on the Knowledge and Modern Language Models

3.1 General Approach to Automated Structuring of Event Information

The proposed approach is aimed at automating the full cycle of processing information about events contained in unstructured textual sources. Unlike solutions focused only on extracting individual facts, it covers retrieval and preparation of relevant texts, extraction of events and their attributes, structuring and normalization of the obtained results, and preparation of the event log representation suitable for the subsequent process mining.

A key feature of the approach is the combination of knowledge-driven and AI-based components. Knowledge-driven components describe the domain, information sources, query history, and processing results, while generative models are used to process variable unstructured texts, identify events, and convert the obtained information into a predefined structure. This combination increases the degree of automation while preserving semantic consistency and enabling reuse of the obtained results.

The general workflow of the proposed approach is shown in Fig. 1.

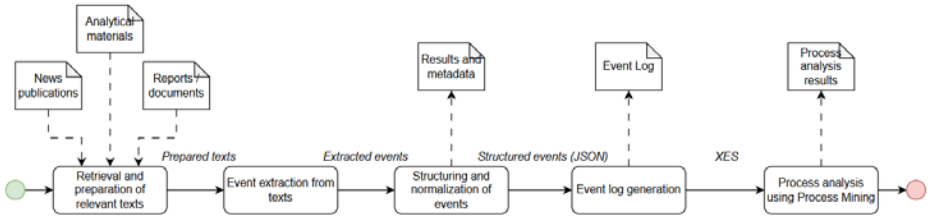

Fig. 1. General workflow of the proposed approach.

The processing starts with retrieval unstructured textual sources and includes finding and preparation relevant texts. The next step is extracting information on events from found and preprocessed text. Then structuring and normalizing extracted information on events are executed. Extracted normalized data presented in JSON format is passed to the next step. Preservation of texts preprocessing results and metadata, event log generated in the XES format for subsequent Process Mining analysis is the last step.

3.2 Generalized Architecture of the Proposed Solution

The general architecture of the proposed solution includes subsystems for text acquisition, text processing using generative models, structuring and normalization of event information, and event log generation in the XES format. The architecture is based on the combination of AI-based and knowledge-driven components and is aimed at supporting the full processing cycle, from text acquisition to preparation of results for subsequent analysis in Process Mining tools. The general architecture of the proposed solution is shown in Fig. 2.

The prototype implementation relies on a modular software organization. Intermediate results are represented in a structured JSON format, which simplifies interaction between the retrieval, extraction, normalization, and events log generation modules. The solution is implemented as a web-based system with a browser interface and backend services responsible for text processing, interaction with generative models through an external API, and transformation of the obtained results into the XES format. Such a design supports extensibility and makes it possible to evolve individual components independently.

The text acquisition subsystem provides access to external textual sources and transfers the prepared text to the subsequent processing stages. The subsystem based on generative models is responsible for retrieving relevant information, extracting events and their attributes, and providing an initial structured representation of the result. The event structuring and normalization subsystem transforms the extracted data into a consistent internal representation required for subsequent events log generation.

An essential role in architecture is played by the ontology-based knowledge layer, which includes the domain ontology and ontology of data sources as well as the ontology of user's queries and results of their execution. This layer is used to define the processing context, support result interpretation, and preserve the interaction history. At the final stage, the XES generation subsystem produces an event log that can be used in process analysis tools, in particular ProM.

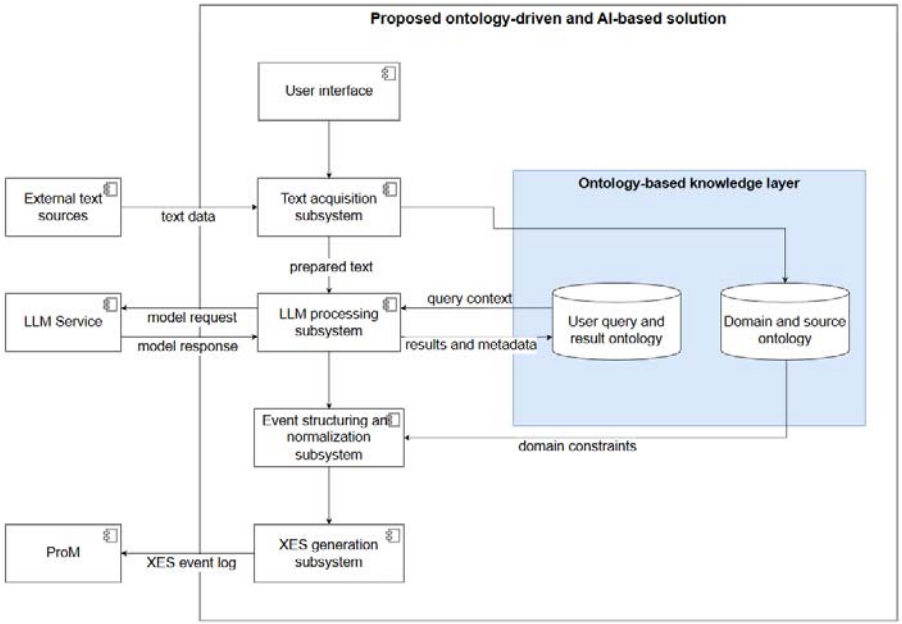

Fig. 2. Generalized architecture of the proposed ontology-driven and AI-based solution.

3.3 Ontological Basis of the Proposed Approach

An essential part of the proposed approach is ontology-based knowledge representation used for retrieval of information on events and preservation of searching and processing results. In accordance with the tasks of the paper, the ontological model can be divided into two main parts: the ontology of the domain and information sources, and the ontology of user interaction and processing results [12-13, 16-20].

The first part describes domains, event types, their attributes, source characteristics, and relations between the main entities (Fig. 3). It formalizes the problem context, extends retrieval through semantic relations between concepts, and supports a consistent interpretation of extracted event information, which is especially important for heterogeneous textual sources, including Internet materials and news publications [10-13, 18-20].

The second part represents interaction between user and system. User's tasks, query formulations, prompts, model invocation parameters, query execution results, and structured representations of extracted events, thus supporting interaction history, result provenance and reuse of previously obtained data are stored in this part of ontology (Fig. 4, 5) [20].

The ontology-based model can be decomposed into several interconnected components supporting different processing stages. Their main content and roles are summarized in Table 2. Such a decomposition makes it possible to connect retrieval, interpretation, preservation of intermediate results, and preparation of structured representations within a unified semantic framework. Together, these two parts form a unified semantic layer connecting source texts, processing parameters, extraction results, and their further transformation into a structure suitable for event log generation in the XES format. Although ontology development remains a non-trivial task, existing ontology-driven approaches show that such models can serve as the basis for intelligent analytical tools supporting retrieval, interpretation, and preservation of processing results [12-13, 19-20].

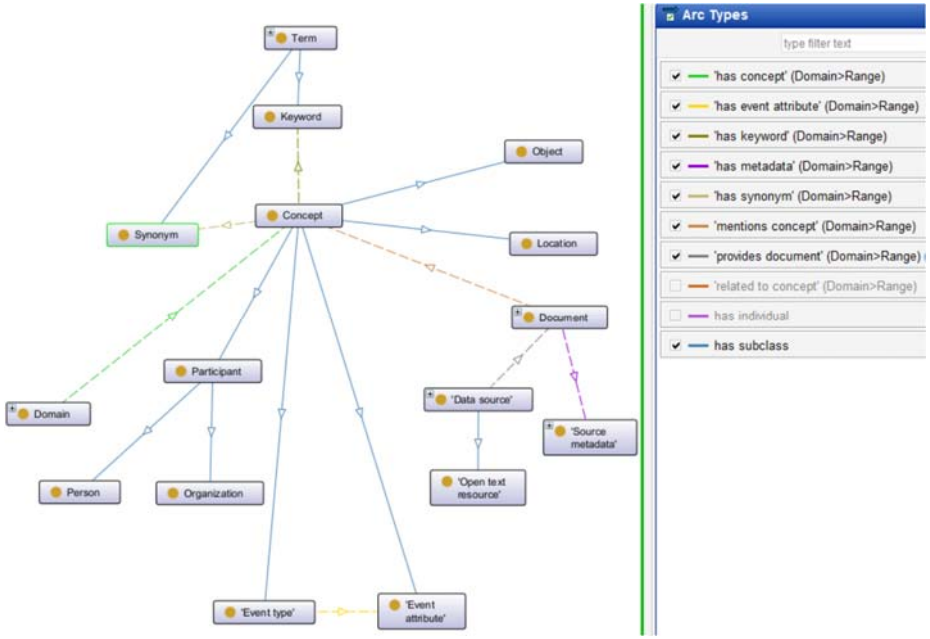

Fig. 3. Fragment of the ontograph of the domain and data information ontology.

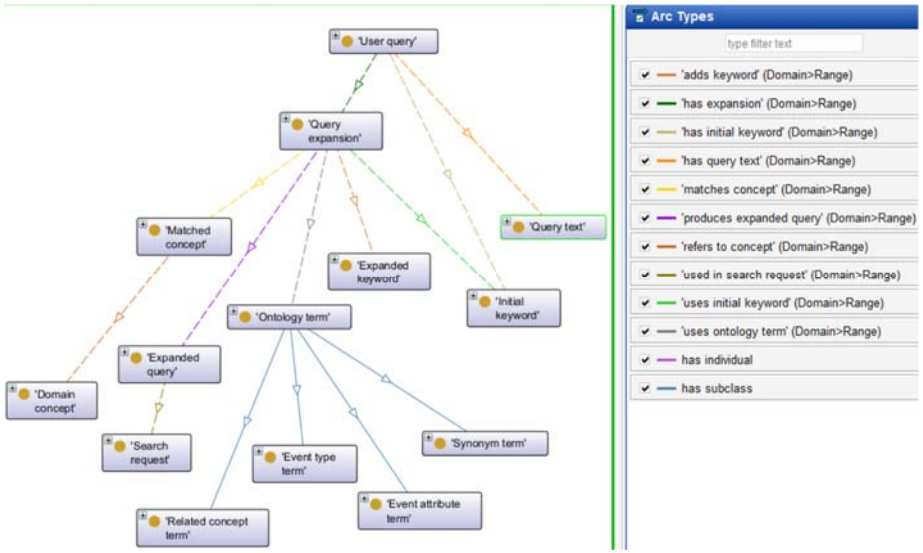

Fig. 4. Fragment of the ontograph of the user query ontological extension.

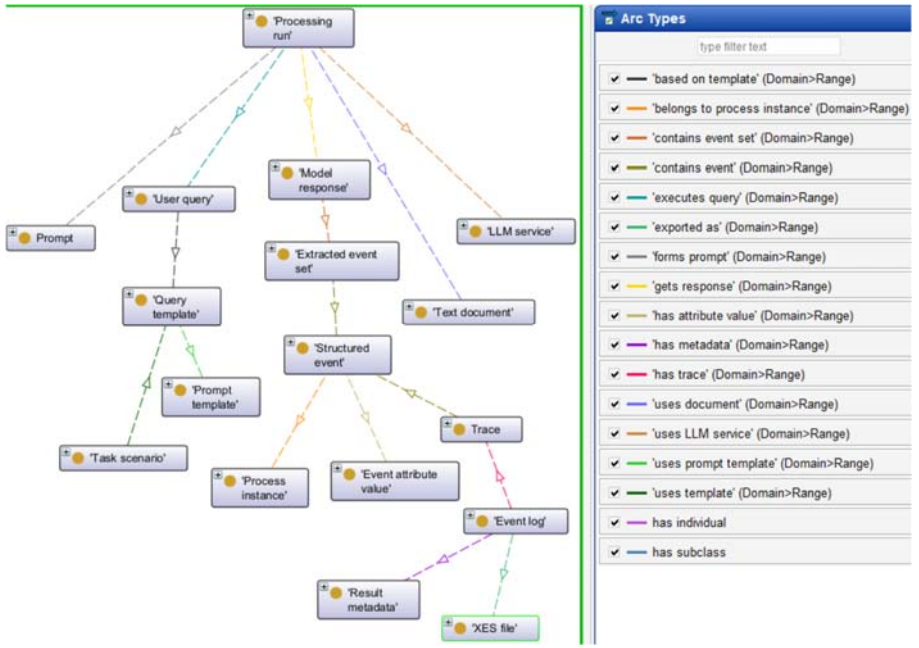

Fig. 5. Fragment of the ontograph of processing results and event log generation.

Table 2. Ontology-based components and their role in the solution.

Component	Stored entities	Role in the solution
Domain model	domain concepts, event types, actors, temporal and spatial attributes	provides semantic context for extraction and normalization
Source model	source types, documents, URLs, source metadata	supports retrieval and source tracking
Query model	user tasks, search queries, prompts, model parameters	preserves interaction history and processing settings
Result model	extracted events, structured JSON representations, metadata	supports result preservation and transformation into XES

3.4 Use of Generative Models in the Proposed Approach

In the proposed approach generative models are used both for retrieving relevant information and for processing unstructured textual sources and transforming extracted information into a predefined structure. Their role includes text retrieval, extraction of events and their attributes, and initial representation of results in a form suitable for subsequent structuring [14-15].

Generative models are especially useful for variable texts in which events, participants, and relations are expressed implicitly or heterogeneously. In such cases, a language model can take context into account, interpret formulations, and produce a more coherent representation of event information. At the same time, the correctness of the result depends substantially on query formulation, completeness of context, and interaction parameters [14-15, 21-22].

When generative models are used at the retrieval stage, irrelevant results and hallucinations are possible. Within the proposed approach, the final decision on the relevance of the retrieved

information remains with the analyst, who may either exclude such results from further processing or consider them as potentially meaningful outliers [14-15].

The generative model is used in combination with the ontology-based foundation of the system. The ontological layer provides the processing context, supports result interpretation, and preserves the history of user queries, prompts, and obtained structured representations [12-13, 20]. Thus, generative models provide flexibility and automation, whereas the ontological basis ensures consistency, traceability, and reuse of results.

To support the user in setting up tasks for searching, extracting and structuring event information, the designed system provides for the use of query templates. The query template in this work is understood as a generalized structure for accessing a system or a generative model, which defines the type of problem to be solved, a set of mandatory and additional parameters, processing restrictions and the expected format of the result. The need to use templates is due to the fact that the result of the texts processing largely depends on the accuracy of the request wording. The same text can be processed differently depending on which events need to be highlighted, which attributes are considered mandatory, which additional characteristics to take into account and in what form the result should be presented. Query templates are therefore seen as a means to unify user scenarios and improve processing reproducibility. In the system being designed, query templates are associated with an ontological layer. Domain and source ontology can be used to substitute keywords, synonymous formulations, related concepts, event types, possible event attributes, and valid sources. The user query extension ontology describes a mechanism for matching source keywords with ontology concepts and generating an extended search query. The ontology of user queries and results, in turn, allows you to save information about the applied templates, their filling parameters, execution results and generated event logs. This creates the basis for reusing successful processing scenarios in similar tasks. Query templates can be used at different stages of the system operation: when searching for text materials, expanding the query ontologically, extracting events, highlighting additional attributes, structuring the result, generating an event log and saving information about the processing performed. In this case, it is advisable to divide the template parameters into several groups (Table 3).

Table 3. Ontology-based components and their role in the solution.

Template type	Template purpose	Template main parameters
Search for text materials	Receive or select source texts for processing	subject, text retrieval method, URL, source type, period, language, keywords, search restrictions
Ontological query extension	Refine original query with ontology terms	source query, keywords, synonyms, related concepts, event types, event attributes
Event retrieval	Highlight basic events in text	text, subject, required event fields, required result format
Advanced event retrieval	Highlight additional event attributes (characteristics)	additional attributes, participants, location, cause, effect, source, metrics, domain constraints
Structuring and normalization	Bringing the result to a single structure	field list, data types, date format, skip rules, normalization rules
Event log generation	Prepare data for XES format	case_id, activity, timestamp, resource, XES extended attributes, grouping rules
Result saving and traceability	Commit completed processing and relationships between objects	запрос, расширенный запрос, источники, результат в формате JSON, события, журнал XES, метаданные обработки

The use of templates also allows you to increase the controllability of interaction with the generative model. In the template, you can predefine the required response format, rules for handling missing values, a list of invalid assumptions, and requirements for the structure of the JSON result. This is especially important when generating event logs automatically, since an incorrect or heterogeneous

model response structure can lead to errors during the normalization and XES file generation stage. Thus, query templates perform several functions in the designed system: set the structure of a user task, define mandatory and additional processing parameters, fix search and retrieval constraints, provide communication with the ontological layer and increase the reproducibility of the result. In combination with the ontological extension of the query, templates allow not only to formalize the user's access to the system, but also to adapt it to the subject area and the requirements of subsequent process analysis.

3.5 Illustrative Workflow of the Proposed Solution

To illustrate the proposed solution, consider a scenario in which an analyst seeks event information about a process contained in news publications. In this case, the system is expected not only to retrieve relevant textual data, but also to transform them into a structure suitable for event log generation.

The first stage of the user script is to search and select text materials containing information about events. This stage is implemented on the Search Texts tab and is designed to form a set of input texts that can later be transferred to the stage of structuring and extracting facts. A feature of this module is the use of a custom ontology to extend the search query. The user can download the domain ontology in .ttl, .rdf, or .owl. After loading, the system processes the file and extracts the main elements from it, including classes, properties, labels, comments, keywords and synonyms. The received data is used to form an ontological context, which is then used to extend the user query. After downloading the ontology, information on its content becomes available to the user. The interface displays the ontological context and provides a visual representation of the ontology.

The next step is to set the user's search query. In the current implementation, the query is generated based on keywords entered by the user through the interface. Each keyword is added to the list of search parameters and can be deleted or replaced if necessary. This approach simplifies the task and allows you to form a request in stages, without the need to manually compose expanded text. If the ontology has been preloaded, the system automatically matches the entered keywords with the terms extracted from the ontology and uses the associated concepts to expand the query. The ontology-based layer provides the processing context, including the domain, event types of interest, possible attributes, and source characteristics. On this basis, the system obtains textual materials from external sources, after which a generative model is used to retrieve relevant information, identify events, participants, temporal characteristics, and other attributes, and provide an initial structured representation of the result.

After the input text is generated, the event extraction and structuring step is performed. The transformation of the unstructured text description into a set of facts presented in a structured form and suitable for the subsequent formation of an event log is the next step. In the implemented prototype, event extraction is performed using an artificial intelligence generative model. The prepared text and system instruction specifying the required response structure are passed to the model input. Unlike a free text response, the result must be returned as a JSON array, where each element corresponds to a separate event. The extracted data are then normalized and transformed into a consistent internal representation. At the same time, the ontological layer preserves the query formulation, processing parameters, and obtained results, which supports interaction history and reuse of these data in similar tasks. As a result, a structured representation of event information is produced, in which individual events can be interpreted as elements of a process. At the final stage, these data are transformed into the XES format and can be used in process analysis tools, in particular in ProM. Thus, the proposed solution supports the full processing cycle, from analytical task definition and information retrieval to event log generation for subsequent process analysis.

Additionally, the user can set domain-specific attributes that reflect the features of the analyzed text. For example, for materials about the sanitary and epidemiological situation, such attributes can be the type of infection, the number of cases; for economic messages - an indicator of cost, price change

or volume of supplies; and so on. Additional attributes can be set in two ways. The first method involves manually adding attributes by the user. The second method is based on the use of an LLM service: the user can request attribute suggestions for the source text, after which the model analyzes the material and returns a list of potentially significant characteristics. Received attributes are not automatically included in the extraction scenario. The user views the list and adds only those attributes that match the purpose of the analysis (Fig. 6). The result of event extraction with additional attributes are shown in Fig. 7.

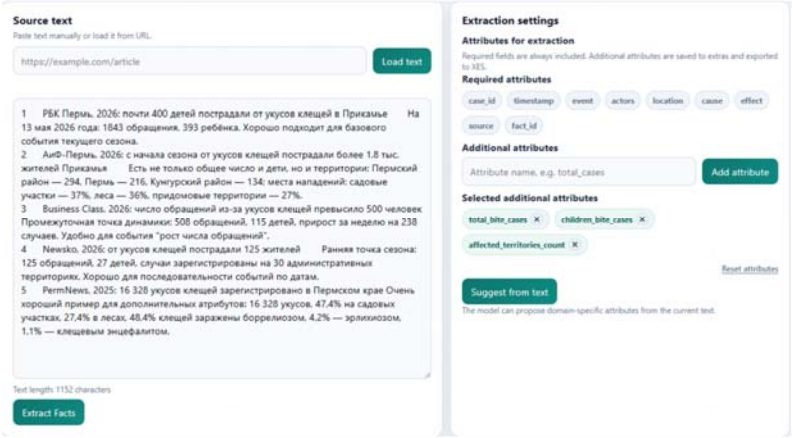

Fig. 6. Specifying source text and additional attributes for event extraction.

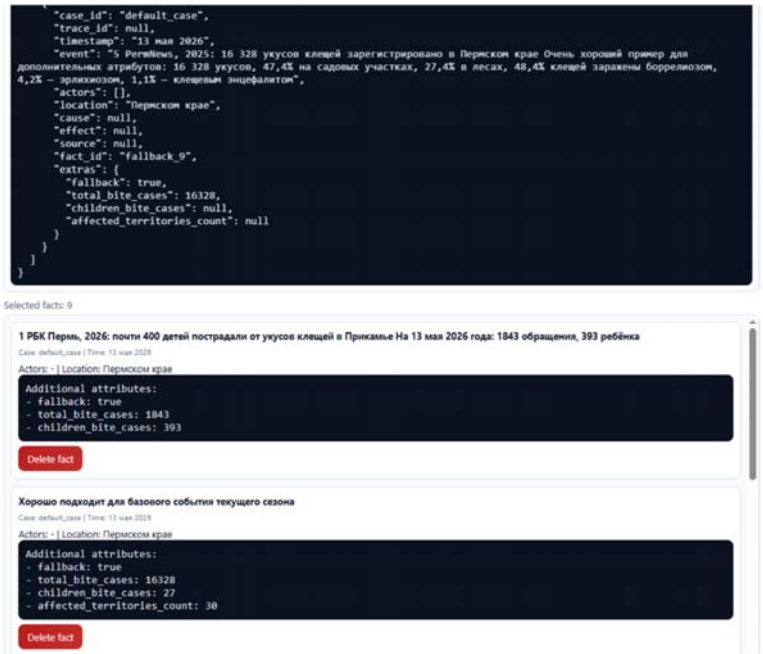

Fig. 7. Presentation of extracted facts with additional attributes.

After extracting facts, the user receives not only a JSON representation of the result, but also a list of facts in the form of separate cards. Each card contains a brief description of the event, process instance ID, timestamp attribute, information on participants and event location if these attributes were highlighted from the text. This presentation makes the result more convenient for viewing and preliminary verification. An important feature of the implemented prototype is the ability to custom select extracted facts. If an individual event was highlighted incorrectly, is not relevant, or should not be included in the resulting event log, the user can delete it before the XES file is generated. This reduces the impact of possible errors in the generative model and improves the quality of the final set of events.

After extraction and fact checking by user, the last step of the processing pipeline is the generation of an event log in XES format. At this step, the structured set of events is converted to the event log format suitable for use in Process Mining tools (Fig. 8).

Fig. 8. Fragment of the generated XES log with additional event attributes.

After successful log generation, the server creates a processing job and stores the result in the internal application store. The link to download the XES file is returned to the user. The generated file can be downloaded through the system interface and then used in external Process Mining tools.

4. Experimental Results and Discussion

Experimental research is aimed at checking the operability of the implemented prototype system for automated generation of event logs based on text data extracted from open Internet sources. The main purpose of the experiment is to assess the possibility of using the developed system to perform a full processing cycle: from the formation of a search query using ontological support to the extraction of facts and the formation of an event log in XES format.

Environmental and man-made accidents were selected as the subject area for the experiment. This area is convenient for checking the developed approach, since information about such events is often

presented in open text sources and contains heterogeneous characteristics: date and place of the incident, participants, causes, consequences, information about the victims, the scale of damage, pollutants and other additional parameters. The presence of such characteristics allows us to check not only the extraction of basic events, but also the possibility of highlighting additional attributes that are significant for subsequent analysis.

The overall experimental study scenario includes the following steps:

1) loading domain user ontology;
2) defining initial keywords;
3) formation of an extended search query using ontology;
4) search for texts according to user's request;
5) selection of relevant materials by user;
6) generation of input text for structuring;
7) defining the additional attributes for extraction;
8) fact extraction using a generative model;
9) verification and selection of extracted facts by user;
10) generation of event log in XES format;
11) analysis of the results.

An experiment is considered successful if the system expands the search query based on ontology, generates a set of relevant text materials, extracts structured facts from them with mandatory and additional attributes, and generates an event log in XES format. Additionally, the possibility of user control at the stages of selecting materials and checking the extracted facts is evaluated, since these operations can reduce the impact of automatic processing errors and improve the quality of the final event log.

To conduct the experiment, a user ontology of the domain of environmental disasters was used (Fig. 9). An ontology contains concepts, keywords, synonyms, related terms, and event attributes that can be used when expanding a search query and setting a subject processing context.

The expected result of using ontology at the query generation stage is the extension of the original user query due to terms associated with the entered keywords. For example, when entering the keyword “accident” the system considers not only direct coincidence with this word, but also synonymous and related concepts. The result of generating a search query using ontological extension is shown in Fig. 10. Some keywords added with ontology were removed to search for events related only to accidents, emissions, oil pollution, and damage.

As a result of the search, the system generated a set of text materials related to environmental and man-made incidents accompanied by a spill of oil, fuel or other petroleum products. Information on events found in Internet sources is presented in the form of cards containing the title, source and a brief text description of the event (Fig. 11). The formed sample included reports of a diesel fuel spill in Norilsk, a tanker accident off the coast of Mauritius, an incident on the Druzhba oil pipeline in the Bryansk region, a fuel oil spill in the port of Nakhodka and an accident at a field in Komi, and so on.

Fact extraction was performed using two generative models connected to the implemented prototype. This made it possible to compare the quality of structuring the same Russian-language input text with the same set of mandatory and additional attributes. The prepared text block, system instruction, mandatory event structure and a list of selected additional attributes: *substance_type*, *substance_volume* and *environmental_objects* were passed to the input of each model.

In the first embodiment, processing was performed using the deepseek-v32 model. This model highlighted a larger number of events, but the result was less suitable for testing work with additional attributes. The model fragmented the source materials into individual events, for example, it

highlighted not only the incident itself, but also individual information about the leak, the participants in the liquidation and the place of pollution. At the same time, additional attributes were not filled in with most cards. This shows that the model was able to select mandatory event attributes but did not fulfill the requirement to extract domain-specific attributes into a given structure.

```
:accident a :Concept ;
    rdfs:label "авария"@ru ;
    :hasKeyword "авария" ;
    :hasSynonym "инцидент", "происшествие", "ЧП", "нештатная ситуация" ;
    :relatedConcept :incident, :emission, :leakage, :pollution, :cleanup .

:emission a :Concept ;
    rdfs:label "выброс"@ru ;
    :hasKeyword "выброс" ;
    :hasSynonym "выброс загрязняющих веществ",
                "выброс вредных веществ",
                "попадание вещества в окружающую среду" ;
    :relatedConcept :pollution, :accident, :chemical_spill .

:pollution a :Concept ;
    rdfs:label "загрязнение"@ru ;
    :hasKeyword "загрязнение" ;
    :hasSynonym "заражение",
                "ухудшение состояния окружающей среды",
                "попадание загрязняющих веществ" ;
    :relatedConcept :emission, :leakage, :waste_discharge,
                    :radiation_contamination .

:evacuation a :Concept ;
    rdfs:label "эвакуация"@ru ;
    :hasKeyword "эвакуация" ;
    :hasSynonym "вывоз людей", "отселение",
                "перемещение населения", "временное размещение" ;
    :relatedConcept :rescue_operation_concept, :emergency_mode_concept .
```

Fig. 9. Fragment of the domain ontology.

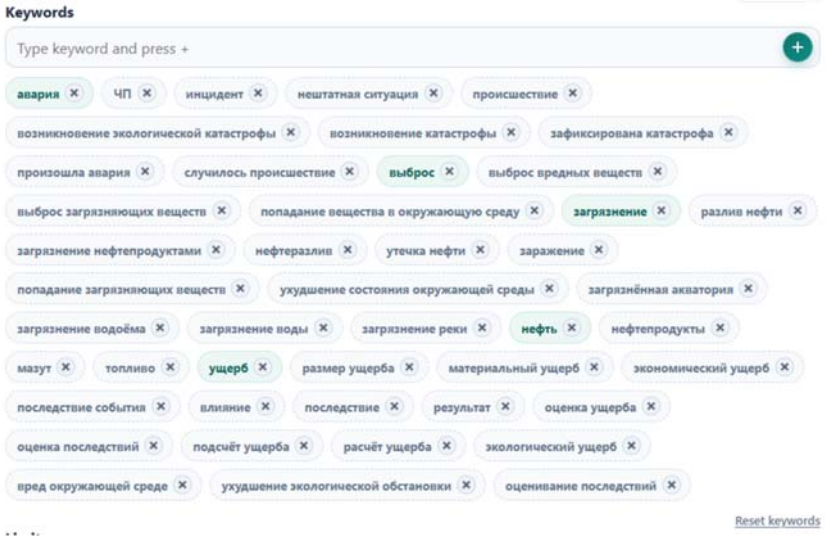

Fig. 10. The result of search query extension with domain ontology.

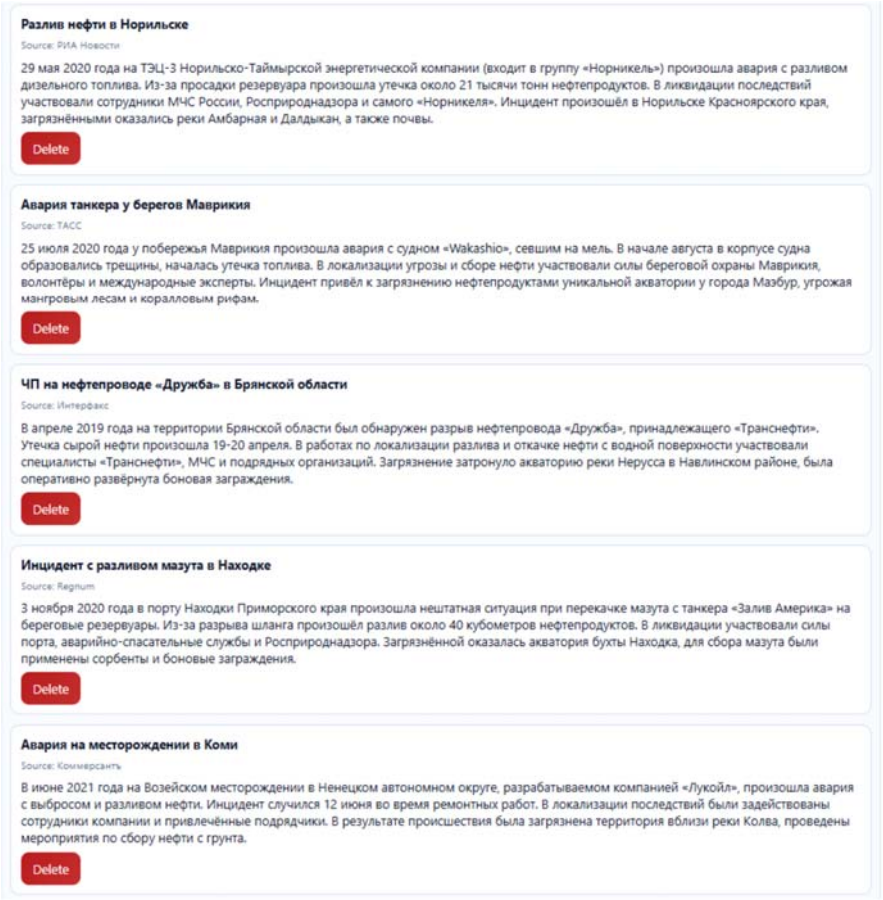

Fig. 11. Search results for text materials with advanced query: event descriptions in cards.

In the second embodiment, processing was performed using the YandexGPT model. This model highlighted fewer events, but the result turned out to be more consistent with the task of the experiment. The events were presented in a more enlarged manner, without excessive fragmentation of the source text into separate fragments. In addition, the model correctly filled in additional attributes: the type of pollutant, the volume of the spill and the affected natural objects were highlighted. For example, for the accident in Norilsk, *substance_type*, *substance_volume* and *environmental_objects* were filled in, which allows you to use the result to demonstrate the extended structure of the event.

The user interface for selecting additional attributes for received texts is shown in Fig. 12.

Comparing the results showed that a higher number of extracted events does not always mean a higher quality result for the event log generation task. For the experiment in question, the consistency of the event structure, filling in the selected additional attributes and the suitability of the result for subsequent conversion to XES format are more important. Therefore, the result obtained using the YandexGPT model was selected for further analysis and event log generation. Log generation results in XES format are shown in Fig. 13.

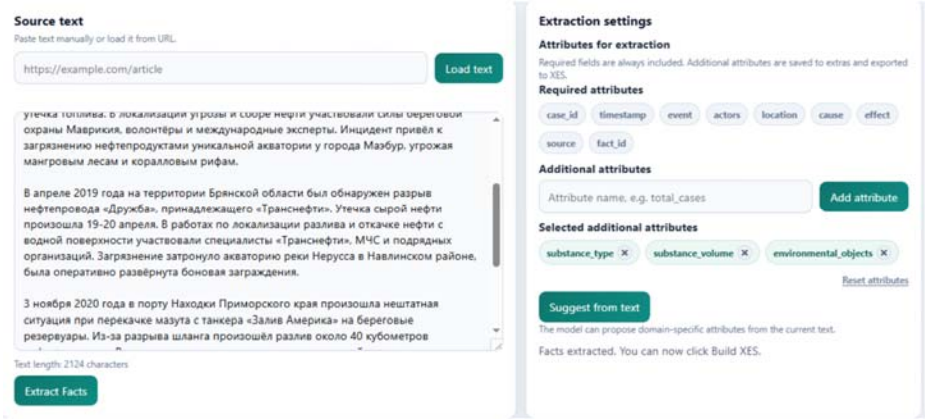

Fig. 12. The user interface for selecting additional attributes for texts containing event descriptions.

```
▾<event>
    <string key="concept:name" value="авария с разливом дизельного топлива"/>
    <date key="time:timestamp" value="2020-05-29T00:00:00"/>
    <string key="location" value="Норильск, Красноярский край"/>
    <string key="cause" value="просадка резервуара"/>
    <string key="effect" value="утечка около 21 тысячи тонн нефтепродуктов, загрязнение рек
    Амбарная и Далдыкан, почв"/>
    <string key="source" value="ТЭЦ-3 Норильско-Таймырской энергетической компании"/>
    <string key="actors" value="сотрудники МЧС России, Росприроднадзор, Норникель"/>
    <string key="fact_id" value="1"/>
    <string key="extra:substance_type" value="дизельное топливо"/>
    <string key="extra:substance_volume" value="21000 тонн"/>
    <string key="extra:environmental_objects" value="реки Амбарная и Далдыкан, почвы"/>
  </event>
▾<event>
    <string key="concept:name" value="обнаружен разрыв нефтепровода «Дружба»"/>
    <date key="time:timestamp" value="2020-04-19T00:00:00"/>
    <string key="location" value="нефтепровод «Дружба»"/>
    <string key="cause" value="разрыв"/>
    <string key="effect" value="утечка сырой нефти"/>
    <string key="source" value="Транснефть"/>
    <string key="actors" value="Транснефть"/>
    <string key="fact_id" value="1"/>
    <string key="extra:substance_type" value="сырая нефть"/>
    <string key="extra:substance_volume" value="null"/>
    <string key="extra:environmental_objects" value="река Нерусса"/>
  </event>
```

Fig. 13. Fragment of the generated XES log with additional event attributes.

The results of the experiment showed that the ontological extension of the request works in accordance with the task. The system is not limited to the user's original keywords, and supplements them with terms related to the selected subject area. To advanced search the context included concepts relating to accidents, oil spills, pollution of water bodies, petroleum products, damage and consequences incidents. This made it possible to obtain materials containing non-general descriptions subject area, and specific event information. The results of the experiment showed that the ontological extension of the request works in accordance with the task. The system is not limited to the user's original keywords, and supplements them with terms related to the selected subject area. This made it possible to extend results, obtain materials containing non-general descriptions subject area, and specific information on events.

5. Conclusion

The paper proposed an approach to the automated search, extraction and structuring of event information from unstructured textual sources for process analysis tasks. It was shown that the combination of knowledge-driven and AI-based components makes it possible to combine semantic consistency, interpretation support, and result preservation with flexibility in retrieval, extraction, and initial structuring of information.

The experimental scenario confirmed the feasibility of retrieving relevant textual materials, extracting events and their attributes, and forming a structured representation suitable for further processing. At the same time, the results showed the need for additional data normalization and expert interpretation of part of the extracted information. Future work is related to the implementation of software in accordance with the proposed architecture, further development of the ontological basis of the solution, and extension of automatic event log generation with tools for processing additional attributes at process analysis.

References

[1]. Marin-Castro H.M., Tello-Leal E. Event Log Preprocessing for Process Mining: A Review. Applied Sciences, 2021, vol. 11, no. 22, article 10556. DOI: 10.3390/app112210556.

[2]. Pegoraro M., van der Aalst W.M.P. Mining Uncertain Event Data in Process Mining. Proc. of the International Conference on Process Mining, 2019, pp. 89-96. DOI: 10.1109/ICPM.2019.00023.

[3]. Alves de Medeiros, A.K., van der Aalst W.M.P. Process Mining towards Semantics. Lecture Notes in Computer Science (LNISA), vol. 4891, 2008, pp. 35-80. DOI: 10.1007/978-3-540-89784-2_3.

[4]. Cremerius J., Weske M. Data-Enhanced Process Models in Process Mining. Preprint. 2021. DOI: 10.48550/arXiv.2107.00565.

[5]. Cremerius J., Weske M. Change Detection in Dynamic Event Attributes. Business Process Management Forum (BPM 2022). LNBIP, 2022, vol. 458, pp. 157-172. DOI: 10.1007/978-3-031-16171-1_10.

[6]. Cremerius J., Weske M. Context-Aware Change Pattern Detection in Event Attributes of Recurring Activities. Intelligent Information Systems (CAiSE 2023). Lecture Notes in Business Information Processing, 2023, vol. 477, pp. 1-8. Springer, Cham. DOI: 10.1007/978-3-031-34674-3_1.

[7]. Lyadova L., Platunov A., Lanin V., Zamyatina E., Matta N. An Approach Based on the Multifaceted Ontology to Development of Event Series Processing Tools. Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023). Communications in Computer and Information Science, 2025, vol. 2454, pp. 232–250. Springer, Cham. DOI: 10.1007/978-3-031-87569-4_11.

[8]. Platunov A., Lyadova L., Matta N., Lanin V., Zamyatina E. An Approach to Developing Ontology-Based Tools for Event Series Analysis. Proc. of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 2: KEOD, 2023, pp. 323-330. DOI: 10.5220/0012259100003598.

[9]. Adnan K., Akbar R. An Analytical Study of Information Extraction from Unstructured and Multidimensional Big Data. Journal of Big Data, vol. 6, issue 1, 2019, pp. 1-12. DOI: 10.1186/s40537-019-0254-8.

[10]. Shalyaeva I., Lyadova L., Lanin V. Events Analysis Based on Internet Information Retrieval and Process Mining Tools. Proc. of the IEEE 10th International Conference on Application of Information and Communication Technologies (AICT 2016), 2016, pp. 168-172. DOI: 10.1109/ICAICT.2016.7991678.

[11]. Shalyaeva I., Lyadova L., Lanin V. Ontology-Driven System for Monitoring Global Processes on Basis of Internet News. Proc. of the IEEE 11th International Conference on Application of Information and Communication Technologies (AICT 2017), 2017, pp. 385–389. DOI: 10.1109/ICAICT.2017.8687086.

[12]. Abrosimova P., Shalyaeva I., Lyadova L. The Ontology-Based Event Mining Tools for Monitoring Global Processes. Proc. of the IEEE 12th International Conference on Application of Information and Communication Technologies (AICT 2018), 2018, pp. 108-113. DOI: 10.1109/ICAICT.2018.8747094.

[13]. Zayakin V.S., Lyadova L.N., Lanin V.V., Zamyatina E.B., Rabchevskiy E.A. An Ontology-Driven Approach to the Analytical Platform Development for Data-Intensive Domains. Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021). Communications in Computer and Information Science, vol. 1718, 2023, pp. 129-149. DOI: 10.1007/978-3-031-35924-8_8.

[14]. Liu Y., Li S. AutoIE: An Automated Framework for Information Extraction from Scientific Literature. arXiv preprint, 2024. DOI: 10.48550/arXiv.2401.16672.

[15]. Johnson T., Clark E., Adams M. LLM-AIx: An Open Source Pipeline for Information Extraction from Unstructured Medical Text Based on Privacy-Preserving Large Language Models. medRxiv preprint, 2024. DOI: 10.1101/2024.09.02.24312917v1.

[16]. Михеев А.А., Ярыгина Е.С. Концепция моделирования семантической разметки корпусов текстов на базе событийной онтологии. Russian Linguistic Bulletin, 2024, № 1 (49). DOI: 10.18454/RULB.2024.49.37. / Mikheev A.A., Yarigina Y.S. The Concept of Modelling Semantic Markup of Text Corpora on the Basis of Event Ontology. Russian Linguistic Bulletin, 2024, no. 1 (49). DOI: 10.18454/RULB.2024.49.37 (in Russian).

[17]. Ma M., Wang P., Yang J., Li C. OntoEvent: An Ontology-Based Event Description Language for Semantic Complex Event Processing. Proc. of the 16th International Conference on Web-Age Information Management, 2015, pp. 348-359. DOI: 10.1007/978-3-319-21042-1_38.

[18]. Miheev A. A. Semantic and Event Ontology EventOnto: Basic Terminological Framework, Development Processes and Principles of Subject Unit Classification. International Research Journal, 2024, no. 8 (146). DOI: 10.60797/IRJ.2024.146.26.

[19]. Lanin V., Lyadova L., Zamyatina E., Vostroknutov N. An Ontology-Based Approach to Social Networks Mining. Proc. of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 2: KEOD, 2021, pp. 234-239. DOI: 10.5220/0010716600003064.

[20]. Zayakin V., Lyadova L., Smirnov M., Lanin V., Matta N., Zamyatina E. Event Series Generation and Analysis Based on Multifaceted Ontology. Proc. of the IEEE 16th International Conference on Application of Information and Communication Technologies, 2022 (AICT 2022), 6 p. DOI: 10.1109/AICT55583.2022.10013573.

[21]. Ritter A., Mausam M., Etzioni O., Clark S. Open Domain Event Extraction from Twitter. Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 1104-1112. DOI: 10.1145/2339530.2339704.

[22]. Yao Z., Li Q., Pan L., Wu Q., Cheng J., Liu Q. A Context- and Time-Aware Model for Knowledge Graph Fact Extraction. Proc. of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 865-873. DOI: 10.1145/3447548.3467272.

Информация об авторах / Information about authors

Кристина Дмитриевна ВОРОНОВА – студент магистратуры Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ – Пермь), образовательная программа «Бизнес-информатика». Сфера научных интересов: анализ процессов, обработка текстов, генеративные модели, онтологическое моделирование.

Kristina Dmitrievna VORONOVA – master student at the National Research University “Higher School of Economics” (HSE University, Perm Branch), educational program “Business Informatics”. Research interests: process analysis, text processing, generative models, ontology-based modeling.

Людмила Николаевна ЛЯДОВА – кандидат физико-математических наук, доцент, доцент кафедры информационных технологий в бизнесе Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ – Пермь). Сфера научных интересов: языки моделирования, предметно-ориентированное моделирование, языковые инструментарии, CASE-средства, системы имитационного моделирования.

Lyudmila Nikolaevna LYADOVA – Cand. Sci. (Phys.-Math.) in Computer Science, Associate Professor of the Department of Information Technologies in Business of the National Research University “Higher School of Economics” (HSE University, Perm Branch). Research interests: modeling languages, domain specific modeling, language toolkits, CASE tools, simulation systems.