

# Events Analysis Based on Internet Information Retrieval and Process Mining Tools

Irina Shalyaeva

Business Informatics Department  
National Research University Higher  
School of Economics  
Perm, Russian Federation  
ishalyaeva@bk.ru

Lyudmila Lyadova

Business Informatics Department  
National Research University Higher  
School of Economics  
Perm, Russian Federation  
llyadova@hse.ru

Viacheslav Lanin

Business Informatics Department  
National Research University Higher  
School of Economics  
Perm, Russian Federation  
vlanin@hse.ru

**Abstract** — This paper presents preliminary result of research project, which is aimed to combine ontology information retrieval technology and process mining tools. The ontologies describing both data domains and data sources are used to search news in the Internet and to extract facts. Process Mining tools allows finding regularities, relations between single events or event types to construct formal models of processes which can be used for the next ensuing analysis by experts. An applicability of the approach is studied with example of the environmental technogenic disasters caused with oil spills, and followed events. Ontologies allow adjustment to new domains.

**Index Terms** — fact extraction, event analysis, text mining, process mining, structure-centered information retrieval.

## I. INTRODUCTION

Nowadays information retrieval tools allow finding the publications containing information on events (facts), on the objects and relations, on the time when events occurred and places and so on [9, 10, 11]. The approach allowing to reveal dependences between events (facts), information on which is published in the Internet, is offered in this paper.

The ontologies describing both data domains and data sources are used to search news in the Internet and to extract facts. The domain ontologies allow to establish connection between objects and events, to classify facts, to execute a clustering, etc. on the base of information published in the Internet. The ontologies of data sources describe the information sources in the Internet. The results of information retrieval are structured and stored in data base.

Event logs are formed on base of the stored data in formats used in Process Mining tools. The data is cleared and detailed at preprocessing with ontologies. Process Mining tools allows to find regularities, relations between single events or event types, to construct formal models of processes which can be used for the next ensuing analysis by experts.

At the research prototype development existing tools of information retrieval are used for events finding and logs generating and ProM system is used for process mining.

Process Mining gradually penetrates into the growing number of applications solution. The opportunity, which provides this discipline, to discover, monitor and improve

processes, aside from such evident purpose as obtaining data from software systems, are used for different tasks like tracing and analyzing students' learning habits based on MOOC data, ontology-driven data extraction from databases, in Workflow Management Systems for the healthcare industry. This research demonstrates that combining Process Mining with other disciplines and approaches can provide a variety of interesting and nontrivial results; this promotes a lot of interest on the part of the scientific environment.

## II. DEFINITIONS AND LIMITATIONS OF THE DOMAIN

As already mentioned, in our work we suggest to combine web, text and process mining in order to obtain unevident data patterns in a convenient and accessible graphical representation with the possibility of further model analysis. To analyze opportunities and demonstrate the described approach, technogenic accident subject area has been chosen, namely the events related to the oil spill. To understand what kinds of data can be collect from news feeds as part of this theme, a lot of query results were analyzed. Russian web-media and global search engines, like Google, were considered as data providers.

Further, assuming that the user may have two basic types of information needs:

- 1) gross appearance in the industry, statistics;
- 2) data on the specific event.

Two corresponding types of requests have been analyzed (f.e. the generalized – “oil spill”, and a specific – “oil spill in Sakhalin April 5, 2016”). Obviously, the more general request is executed, the more diverse data we receive. So, by request for oil spill, results can be related to: the elimination of consequences, the sanctions measures, the new methodology to eliminate spills, actually oil disaster and many other types of events.

In this paper, we consider simple referential events:

- N. Samoilenko characterizes referential event as: “by event we understand the result of an action, behavior, occurrence, fact, which has a personal or social significance, something new, a change in the situation, the state of affairs”.
- Simple events – internal form consists of the primary elements of event, first of all – action and associated

components of activity.

The minimal set of characteristics that we takes into account for events analysis are: participants of the event, geographical location, event border, internal relations between components of a single event, the relation between events. Such static event attributes as the company, the date format and geographical position will be a part of the domain ontology. Events borders were identified in the analysis of news feeds texts on the subject of oil spills. Since the news reports generally displayed the most important aspects of the case described, the event is considered as described in the ontology instance of the class “event” or, if there was no match in the ontology, any verbal constructions: verb + related words, met in the summary of the article. Communication components inside the event set during the semantic analysis that defines the verb structure to identify events. Dependency tree is constructed for each verb. Further, there is the extraction of causal relations. This task is also facilitated by the analysis of news reports, since the size of the text is usually limited from two to five sentences, and the sequence of events often correspond to the stacking order. For escalating a chain of events, sequences extracted from various news reports will be linked by a given feature. News title does not belong to processable text, because it can significantly disturb the process of causal links identifying and duplicate information from the news. However, the information from the header can be used to retrieve objects and event attributes if the event is duplicated in the news text. Also, an event in the header can be used as a marker for a situation by which a news-related situation will bind during further processing. In other words, by these markers, we can associate the events to traces. Trace in the Process Mining is a sequence of events, united by a common use case or a news message in our work. The event is an instance of activity that represents a well-defined process step. Traces are a display of the processes interaction.

### III. FILLING THE DATABASE TO GENERATE MODELS

Obviously, on the specific request we get most of the same set of events. If this request is generated for a statistic model and converts to a single trace for the overall process, a problem arises. The problem of information duplication in the traces can be solved by means of Process Mining. However, synonymous in the data must be unified with the use of ontology.

Data retrieved from Internet via ontologies stored in the database in the same way as user queries on which we were looking for news.

If the user has generated a request, which items are not in our ontology, we offer it to expand the ontology and may suggest where he could add the concept. As a result of the work on this step system shows the user settings from the ontology that it can optionally specify (for example, to limit the time period, select a location, district, company). Then the system generates the necessary additional requests for news portals, extracts texts, extracts the information and fills the database.

After all, the user is again asked to select what data parameters for the model he is interested in. For that the

original user request, extracted data and the data stored in database are analyzed. Summarizes some common virtual representation of the data relating to all the parameters of the request, on the basis of which the user is prompted to choose the criteria that he wants to be taken into account for the construction of the process model. Based on this creates RDF file for storage and log generation.

### IV. USING PAGE STRUCTURE ONTOLOGY FOR INFORMATION EXTRACTION

In our approach the information search method based on web documents structure analysis and ontologies implementation is offered. Two-level ontology capturing following description is supposed to be developed:

- website (the web document being analyzed source) structure description – main page types and their interconnections;
- Web page information blocks description and their interconnections.

The example fragment of these two levels is depicted in the Fig. 1. The first level ontology per se keeps the description of pages existing on the web site in question – sitemap, but in more simplified and generic form. While developing a specific website description, ontology nodes will be populated with the addresses of visited and analyzed pages. The second level ontology is aimed at keeping the description of information blocks to be found on a web page, for instance, navigation block, which can contain valuable information, as well as that of forming these blocks like form controls, static or dynamic images, tables, text areas and so on and so forth. In order to develop this second level ontology the most widespread HTML template kinds were examined along with template provision elements. While developing a specific webpage description, ontology nodes are supposed to capture markup places unambiguously identifying exact placements of webpage information blocks for further data extraction from it or from lower-level element constituting the block in question.

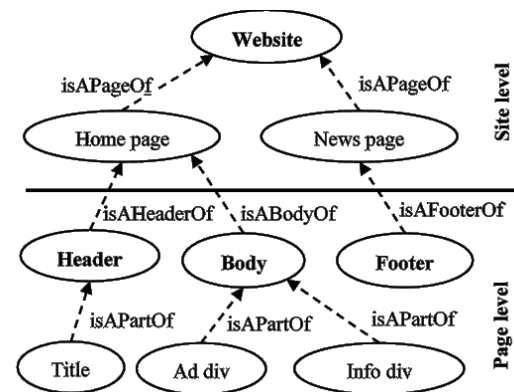


Fig. 1. Example Fragment of Two-level Web Document Ontology

In order to extract information from web documents by means of the proposed two-level structure ontology it is necessary to traverse it and identify the exact information block position for further content query.

Generally, there are several stages in the proposed ontology processing mechanism:

- Loading the ontology from local file or by the URI generated by the ontology editor exploited.
- Traversal algorithm execution and information block identification.
- Saving the placement address of the information division founded for further content query processing.

The main practical application to the offered method is to find more advantageous solution to the information extraction problem by boosting result relevance level paying more attention to the structure and placement of information.

The main advantages to this structure-centered information retrieval approach are as follows:

- web document can be annotated with the structure metadata allowing to take information placement into account;
- structure-centered information retrieval considers and exploits information divisions hierarchical structure interconnections;
- information placement metadata can help identify content duplication and filter it afterwards;
- return result representation is to be enhanced by using structure and placement metadata.

#### V. THE DATA STRUCTURE / DATA CLASSIFICATION / DATATYPING FOR STORING

In processing the results of the above general type of queries, in addition to the described steps for each news item or a trace, it is also necessary to classify the situation described (as shown in Figure 1, the results for "oil spill" are different types of events: social, political and environmental). For each event class should be allocated own attributes that can act as markers depending on the model that the user wants to receive.

Events are classified in terms of the attributes that they may possess. One news post may include events related to different classes. Key attributes are applying to connect them.

The following main types of events and key attributes were identified during the analysis of the news feed of oil disasters:

- Disaster (date, oil company, place) – directly disaster themselves, such as fire, spill, explosion.
- Financial implication (organization) – assessing the financial damage, this includes as a cost for the elimination of consequences as well other economic indicators of enterprises, population and countries.
- Industry news (oil company, publication date) – possible scientific discoveries, achievements, innovations in the field of oil industry, any information related to the operation of companies: enlargement, closure, bankruptcy.
- Sanction (date) – information on the sanctions and penalties.
- Socio-environmental implication (publication date) – the impact on the population, the victims, damage to agriculture, the impact on society and the possible reactions, demonstrations, unrest.

- Socio-political (Date, Place) – influence on government policy, changes in relationships, the impact on foreign trade, sea routes.
- Noise – data from the news that does not belong to a domain but are the results of a query in the Internet search engines (f.e. Define Oil spill at Dictionary.com).

#### VI. MULTIPLE ADVANCED SEARCH QUERIES

As already mentioned, once the data extracted by user request, they are analyzed and, depending on the completeness of the data set the user is able to:

1. Expand the data set (Query for a specific event and query by attributes (geographical position, the company, the time interval)).

2. Build a process models that are available at this stage.

If the user decides to supplement the data with repeated request, the system provides the opportunity to extend the ontology, and using the concept of the updated domain ontology to generate a new request to the network news portals. Advanced data set allows to create more accurate and full covered process models.

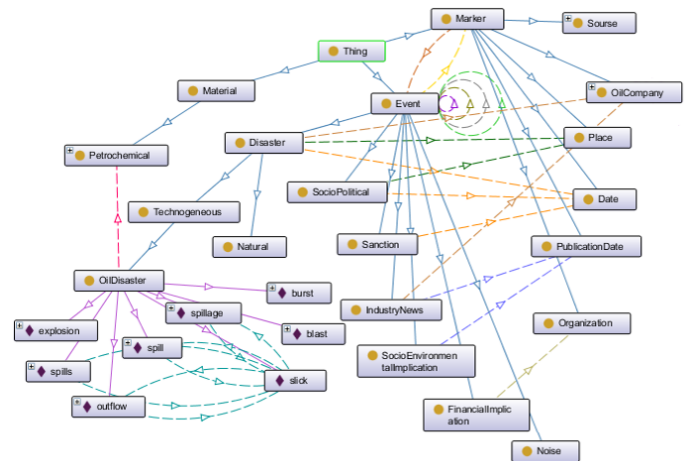


Fig. 2. A fragment of the domain ontology example

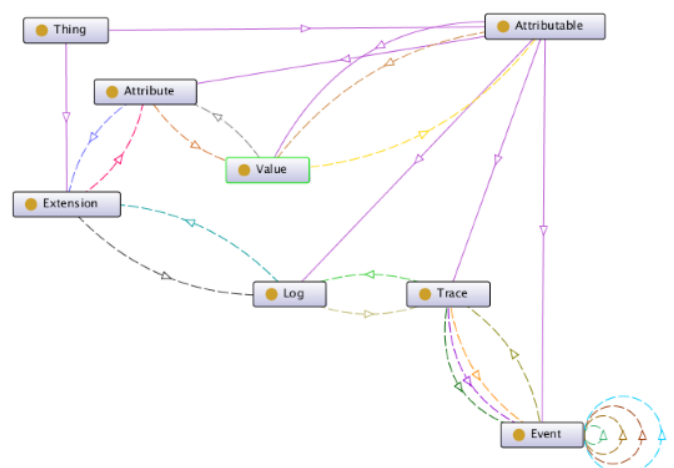


Fig. 3. XES log ontology

Further, the concepts of the domain ontology (Fig. 2) are mapped to the ontology of the log (Fig. 3), and the output document in XES format creates. As a standard set of log extension used to describe data is not enough for our domain, it is necessary to expand. In addition to the below listed standard extensions, we need to include the date of publication, name of the organization that are not directly related to petroleum activities, the location of the event and the source of the news posts.

#### VII. USING THE CAPABILITIES OF STANDARD XES EXTENSIONS

Concept extension – store the name of all levels of the log hierarchy elements: case, trace, event. The instance attribute store the identifier of the case so that identical events may differ from each other.

Lifecycle extension – reflects the stage of process lifecycle, described in the transactional model. This model can be arbitrarily set or use one of the standard models – BPAF or Standard lifecycle transition model.

In our case, it does not matter what kind of lifecycle model to choose because we will not have a hard definition of the beginning and end of the process. Each event by itself is a finite process, and depending on the chain of events, which is found, the start and the end of a whole process can be different in each situation. Therefore, we will not have a standard separation like Running or Not Running process (Fig. 4).

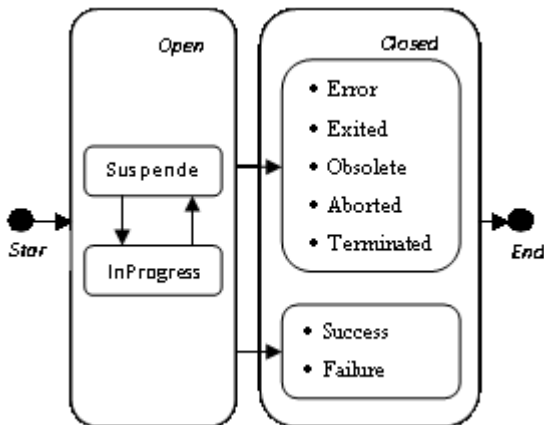


Fig. 4. Lifecycle transactional model

However, it is possible to define some time limit for the determination of completion of various classes of events. For example, if we consider an event that took place 5 years ago, and all related events dated to the same time, the process is considered complete.

If we look at recent events, such as the oil spill occurred yesterday, it is obvious that the response to this event may still appear in the news, thus it is necessary to keep tracking the process. And then if the user wants to know how yesterday's oil spill reflect on the company's present, the model can not be constructed as we do not find relevant news and the process

receives the status COMPLET = failure. In this case, it is also possible to be in state IN PROGRESS, if passed a certain time period f.e. a year after the start of the trial, and not found any effects events – SUSPENDED.

The beginning and the end for the different types of models as determined individually. If the model is typed, then START and END are already set. When the user builds a new unique model, a model of individual events (a specific disaster), the system will build a model by making the assumption that the event which has no predecessors is a START, and the one which has no followers – END.

The user is also given the opportunity to set these parameters of the model if the system mistaken.

Organizational extension – identifies three attributes for the events that identify the actor responsible for the event, and its position in the organizational structure.

Time extension – date and time when the event happened.

Semantic extension defines an attribute that allows to store a number of references to model concepts in a domain ontology in any element of the XES type hierarchy.

ID extension – provides unique identifiers for all elements of log hierarchy.

Cost extension – in our case carries information about the size of fines, eliminating price effects, damage assessment and other monetary indicators [8].

#### VIII. EXAMPLE

According to the data obtained by a general request, the system can build a very simple process models within the same news. Obviously, this is not enough, so the user is prompted to specify a request, choose the information about Shell Spills Oil in the Gulf.

Extraction data processing from Internet news and data processing is executed by means of RapidMiner. Search results are presented in the form of the table containing data on events (facts). The constructed table is transformed into log format. Formal model generated with ProM on the base of this log is given below (Fig. 5).

This model isn't informative:

- Each route is provided as separate option of succession of events.
- Synonymous concepts weren't grouped and brought to a uniform look therefore algorithms of the intellectual analysis of processes couldn't structure and reveal communication in routes and the sequences of events.

It is possible to mark the following shortcomings:

- Absence of a possibility of detection of causes and effect relationships.
- A large number of errors and losses in case of extraction of events.

The revealed problems are fixed when using of the constructed ontologies which fragments are shown above.

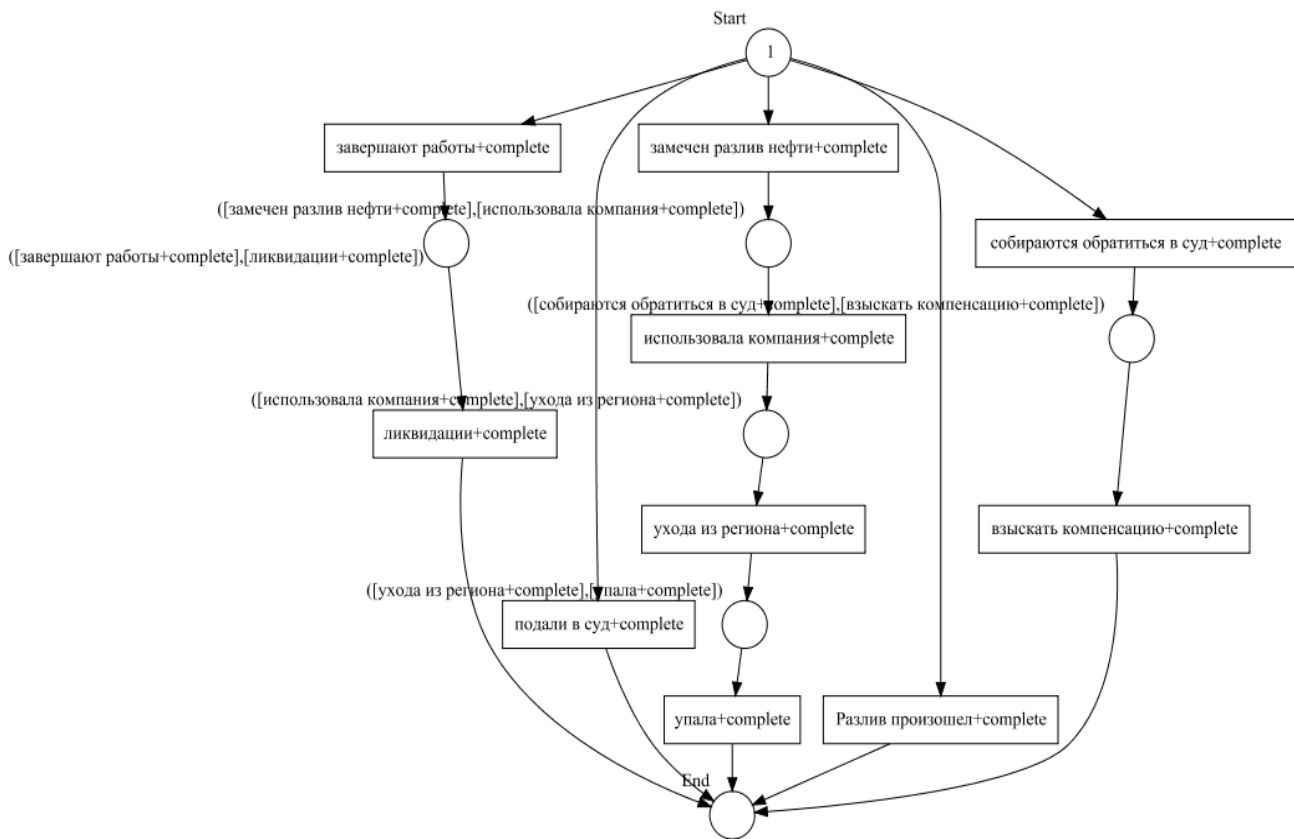


Fig. 5. Formal model generated with ProM on the base of facts table extracted from news

Resulting log generated with using ontologies is given below:

```
<?xml version="1.0" encoding="UTF-8" ?>
<log xes.version="1.0" xes.features="nested-attributes"
openxes.version="1.0RC7" xmlns="http://www.xes-standard.org/">
  <extension name="Lifecycle" prefix="lifecycle"
uri="http://www.xes-standard.org/lifecycle.xesext"/>
  <extension name="Organizational" prefix="org"
uri="http://www.xes-standard.org/org.xesext"/>
  <extension name="Time" prefix="time"
uri="http://www.xes-standard.org/time.xesext"/>
  <extension name="Concept" prefix="concept"
uri="http://www.xes-standard.org/concept.xesext"/>
  <extension name="Semantic" prefix="semantic"
uri="http://www.xes-standard.org/semantic.xesext"/>
  <global scope="trace">
    <string key="concept:name" value="__INVALID__"/>
  </global>
  <global scope="event">
    <string key="concept:name" value="__INVALID__"/>
    <string key="lifecycle:transition" value="complete"/>
  </global>
  <classifier name="MXML Legacy Classifier"
keys="concept:name lifecycle:transition"/>
  <classifier name="Event Name" keys="concept:name"/>
  <classifier name="Resource" keys="org:resource"/>
  <string key="source" value="Rapid Synthesizer"/>
  <string key="concept:name" value="exercice1.mxml"/>
  <string key="lifecycle:model" value="standard"/>
  <trace>
    <string key="concept:name" value="Case3.0"/>
```

```
<event>
  <string key="org:resource" value="Shell"/>
  <date key="time:timestamp" value="2016-05-12"/>
  <string key="concept:name"
value="subsea flow lines sprung a leak"/>
  <string key="lifecycle:transition"
value="complete"/>
</event>
<event>
  <string key="org:resource" value="Shell"/>
  <date key="time:timestamp" value="2016-05-12"/>
  <string key="concept:name"
value="oil spilled into the Gulf of Mexico"/>
  <string key="lifecycle:transition"
value="complete"/>
</event>
<event>
  <string key="org:resource" value="Shell"/>
  <date key="time:timestamp" value="2016-05-19"/>
  <string key="concept:name"
value="cleanup operation ends"/>
  <string key="lifecycle:transition"
value="complete"/>
</event>
</trace>
</log>
```

Petri net model generated according to this log for request “Shell Spills Oil in the Gulf” is shown below (Fig. 6).

Further presented the model of the same process in the context of geographical position, i.e. request was extended by

obtained in the previous step attribute Place = Mexico. The results are shown in Fig. 7.

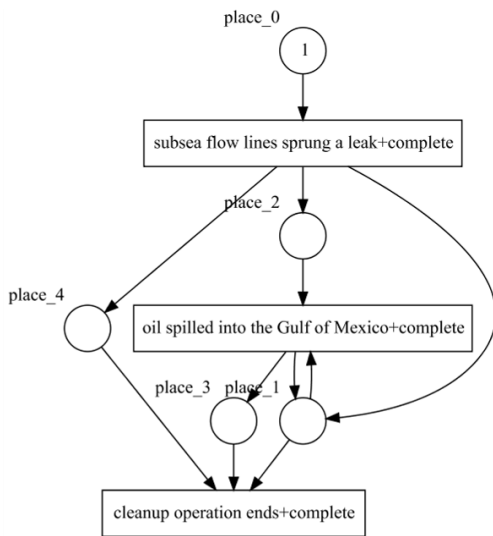


Fig. 6. Petri net model for request “Shell Spills Oil in the Gulf”

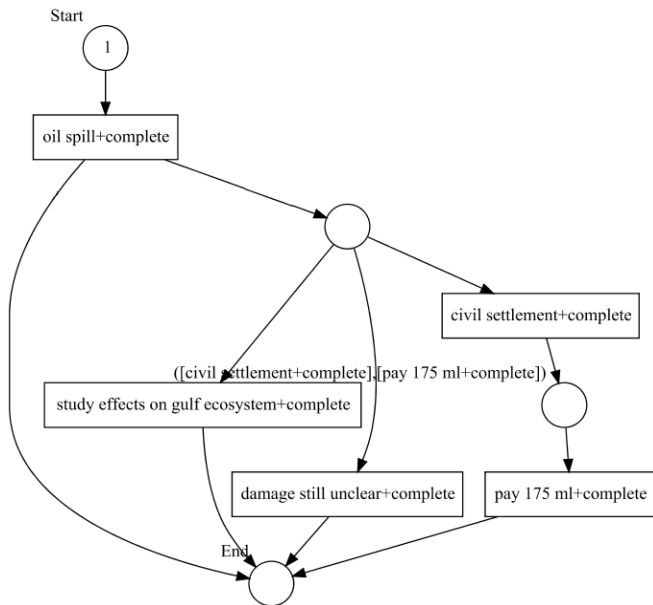


Fig. 7. Model for information request enhanced by parameter Place

In the model appeared some additional information on the socio-economic developments related to the assessment of damage and the imposition of a fine. Thus, clarifying and

expanding the information requests the user fills the database with new data and gets more detailed models.

## IX. CONCLUSION

An approbation of the offered approach for the analysis of the processes associated with the environmental technogenic catastrophes caused with oil spills showed prospects of the developed means. Ontologies allow performing flexible tuning for various domains, for example to investigate relations between events in the field of economy, policy and etc.

## REFERENCES

- [1] M. Leemans; W.M.P. van der Aalst. Process Mining in Software Systems: Discovering Real-Life Business Transactions and Process Models from Distributed Systems // In: Proceedings of 18th International Conference on Model Driven Engineering Languages and Systems (MODELS), 2015. Pp.44-53.
- [2] P. Mukala, J. Buijs, M. Leemans, W. van der Aalst. Learning Analytics on Coursera Event Data: A Process Mining Approach. In: Proceedings 5th International Symposium on Data-driven Process Discovery and Analysis 2015. Pp.18-32.
- [3] D. Calvanese, M. Montali, A. Syamsiyah, W.M.P. van der Aalst. Ontology-Driven Extraction of Event Logs from Relational Databases. In: Business Process Management Workshops 2015.
- [4] R.S. Mans. Workflow Support for the Healthcare Domain. PhD Thesis. Technische Universiteit Eindhoven, Eindhoven, 2011.
- [5] N.A. Samojlenko. Semantika sobytijnosti i sposoby ee vyrazheniya: avtoref. dis. kand. filol. nauk. Alma-Ata, 1991.
- [6] V.E. Goldin. Imena rechevuh sobytij, postupkov i zhanry russkoi rechi // Zhanry Rechi. – Saratov, 1977. – Pp.23-34.
- [7] P.P. Maslov. obnaruzhenie i izvlechenie prichinno-sledstvennyh zakonomernostej iz teksta na estestvennom yazyke. In: Proceedings of Conference “Znaniya-Ontologii-Teorii”. 2009.
- [8] Draft Standard for XES - eXtensible Event Stream - for achieving interoperability in event logs and event streams. The Institute of Electrical and Electronics Engineers, 2016.
- [9] V. Peña-Araya. Galean: Visualization of Geolocated News Events from Social Media / V. Peña-Araya, M. Quezada, B. Poblete // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). ACM New York. 2015. Pp. 1041-1042.
- [10] M. Schuhmacher. Finding Relevant Relations in Relevant Documents. In: Advances in Information Retrieval: Proceedings of 38th European Conference on IR Research, ECIR. Padua, Italy, March 20-23, 2016. Pp. 654-660.
- [11] A. Vokhmintsev, A. Melnikov. The Knowledge on the Basis of Fact Analysis in Business Intelligence. In: Digital Product- and Process Development Systems.– 2013.– Pp.134-141.