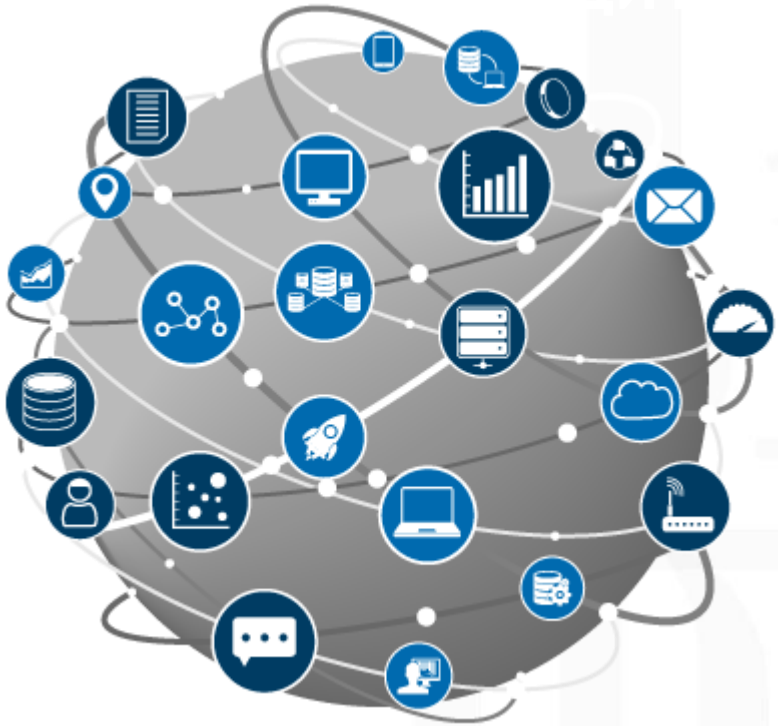
Ирина Шаляева, Вячеслав Ланин, Людмила Лядова

Подход к мониторингу глобальных процессов на основе Интернет-новостей с помощью средств Process Mining

Актуальность

Информационная потребность:

установить связи между событиями, происходящими в мире



Постановка задачи

Какие причины?
Что предшествовало?

Какие последствия?
Что вызвало?

Постановка задачи

Какие причины?
Что предшествовало?

Исследовать возможность построения формальных моделей процессов на основе публикаций о событиях в Интернет с помощью средств поиска и инструментов Process Mining

Что вызвало?

Задачи Process Mining

Особенности глобальных процессов

Особенности глобальных процессов:

- **Границы процессов значительно шире организационных.**
- **Нет структурированных источников данных о событиях.**
- **Пользователи не обладают полными знаниями о процессах (типах событий, объектах, операциях…).**

Process Mining для анализа глобальных процессов

Предметная область

Какие причины?
Что предшествовало?

Какие последствия?
Что вызвало?

Источник данных - новости

Режим ЧС в связи **разливом нефти** введет в районе Коми ...
RusEnergy - 14 апр. 2016 г.
... **разливом нефти** введет в районе Коми, **причины** загрязнения не ... введен в Ухтинском районе Коми, где произошел **разлив нефти**, ...

Водозабору города Ухта угрожает **разлив не**
Гринпис России - 14 апр. 2016 г.
Источник **разлива нефти** пока официально не назва
возможной **причиной** МЧС по Коми считает старые с
1950-х годов, ...

В Ухте из водопроводных кранов может пойти **нефть**. В городе ЧС
Аналитика - Правда.Ру - 14 апр. 2016 г.

Статьи по теме (Ещё 34 статьи)

Структуре «**Роснефти**» выписали **штраф** почти на 7 ...
РосбалтRU - 18 апр. 2016 г.
Суд удовлетворил иск управления Россельхознадзора по Ставропольскому краю к структуре «**Роснефти**» о взыскании

После **взрыва** на заводе Pemex в **Мексике** 18 челове...
РИА Новости - 1 час назад
По последним данным, в результате **взрыва** погибли 13 человек и 136 ... без вести после **взрыва** на заводе **мексиканской нефтяной** компании .... службе по надзору в сфере **связи**, информационных технологий и ...

Взрыв на химзаводе в **Мексике**, десятки пострадавших
Известия - 16 ч. назад

Три человека погибли и более ста пострадали в результате ...
swissinfo.ch - 12 ч. назад

В **Мексике** увеличилось количество жертв в результате **взрыва**
Редакционные - Подробности - 11 ч. назад

В Мексике произошел мощный **взрыв** на нефтяном заводе Pemex
NEWSru.com - 19 ч. назад

еме (Ещё 173 статьи)

Без плана **ликвидации** аварий компаниям запретят добывать ...

Мексика сократит добычу **нефти** на 100 тысяч барре...
Российская Газета - 29 февр. 2016 г.
Сокращение расходной части бюджета **мексиканской нефтяной** компании Pemex на 100 млрд песо (порядка 5,5 млрд долларов) ...

97108

Neftegaz.RU

нефти ...

Минприроды РФ запретит нефтегазовым ком
Neftegaz.RU - 7 апр. 2016 г.

Статьи по теме (Ещё 4 статьи)

Причиной разлива нефти в удмуртской реке стала у...
Interfax Russia - 12 апр. 2016 г.
Interfax-Russia.ru - Специалисты ОАО "Удмуртнефть" (входит в НК "Роснефть" выясняют **причины** утечки нефтесодержащей жидкости в ...

В Удмуртии устраняют последствия **разлива нефти**
ИА REGNUM - 11 апр. 2016 г.

Статьи по теме (Ещё 48 статей)

Общая схема: этап 1

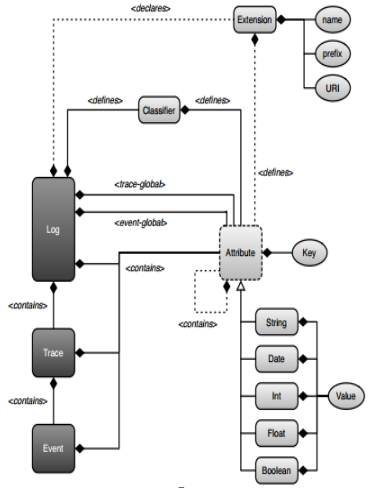
Система поиска информации

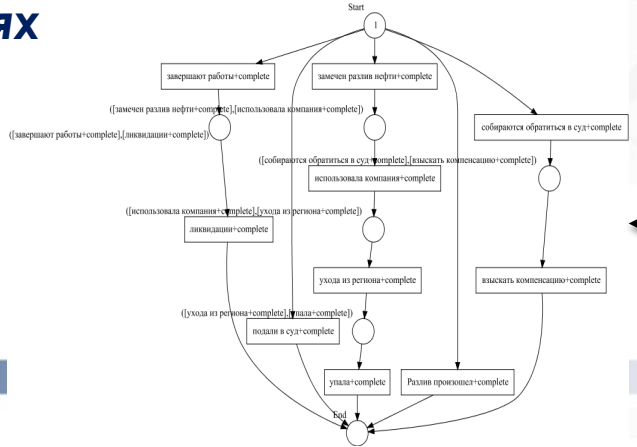
Тождество	Дуальные	Родственные	Полу-дуальные	Заказ
Дуальные	Тождество	Полу-дуальные	Родственные	Ревизия
Родственные	Полу-дуальные	Тождество	Дуальные	Активация
Полу-дуальные	Родственные	Дуальные	Тождество	Зеркальные
Заказ	Ревизия	Активация	Зеркальные	Тождество
Ревизия	Заказ	Зеркальные	Активация	Дуальные
Квази-тождество	Конфликтные	Заказ	Ревизия	Родственные
Конфликтные	Квази-	Ревизия	Заказ	Полу-

Табличное представление данных о событиях

?

Лог событий





Общая схема: этап 2

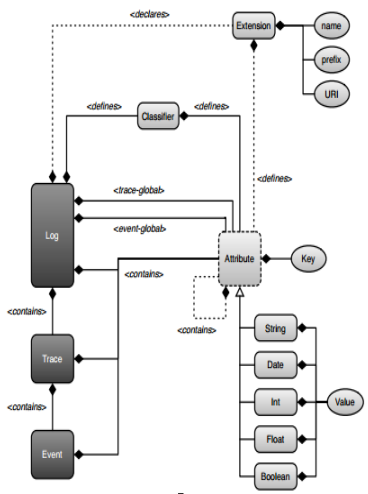
Система поиска информации

БЗ

База новостей

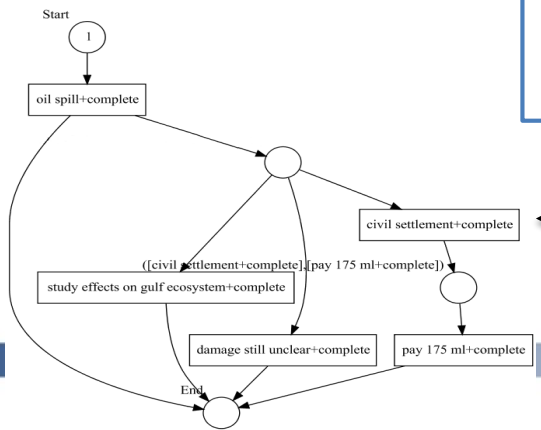
Система подготовки данных

Логи событий



Тождество	Дуальные	Родственные	Полу-дуальные	Заказ
Дуальные	Тождество	Полу-дуальные	Родственные	Ревизия
Родственные	Полу-дуальные	Тождество	Дуальные	Активация
Полу-дуальные	Родственные	Дуальные	Тождество	Зеркальные
Заказ	Ревизия	Активация	Зеркальные	Тождество
Ревизия	Заказ	Зеркальные	Активация	Дуальные
Квази-тождество	Конфликтные	Заказ	Ревизия	Родственные
Конфликтные	Квази-	Ревизия	Заказ	Полу-

Табличное представление данных о событиях



Анализ инструментов информационного поиска

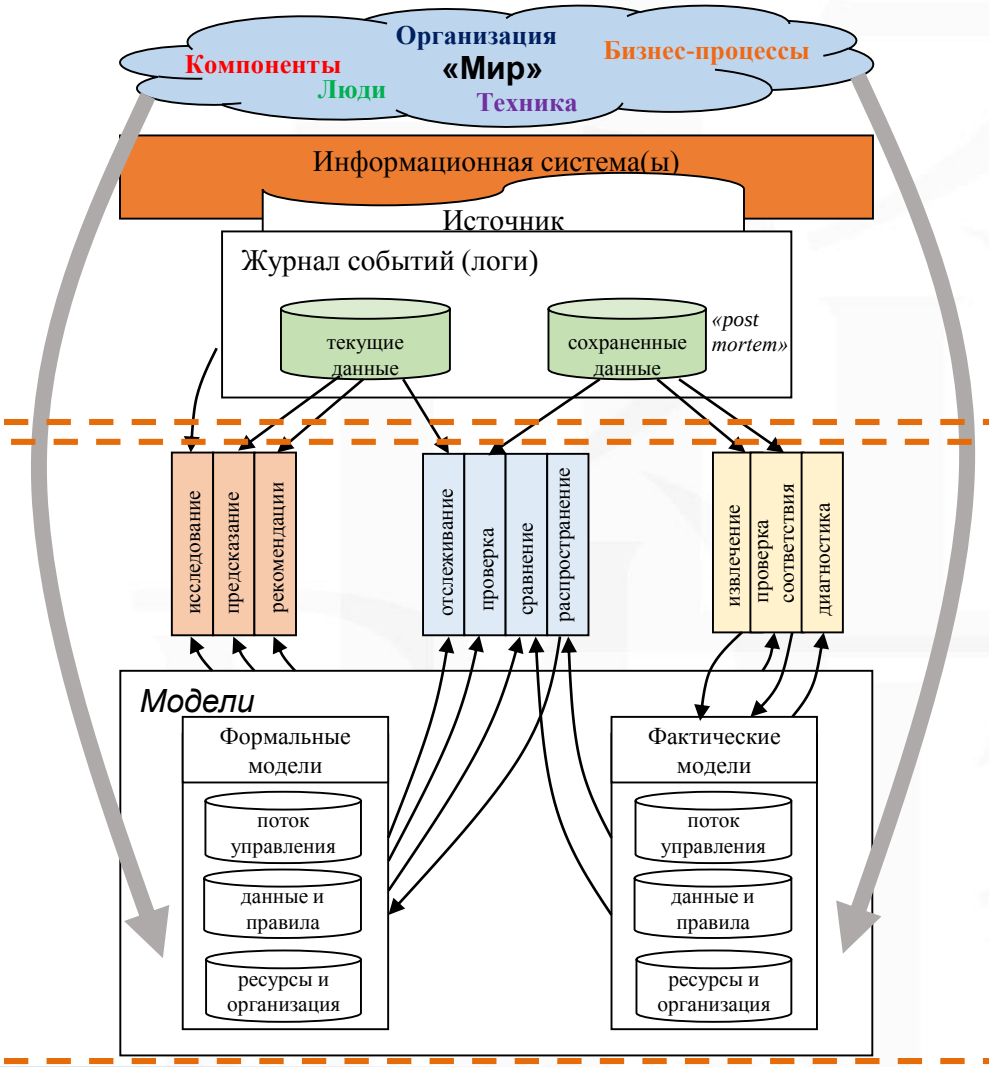
		Google news	Yandex news	Новостные агрегаторы	Rapid Miner	GATE
Поиск по запросу		+	+	+		
Автоматическое аннотирование		+	+	+		
Классификация и кластеризация					+	+
Задача извлечения фактов:	Организации				+	+
	Даты				+	+
	Источники сообщений				+	+
	Географические объекты				+	+
	События				-/+	-/+
Работать с русскоязычными текстами					-/+	-/+
Извлечение веб-ресурсов					+	+
Экспорт результатов					+	-
Доступность системы					+	+
Возможность расширения системы					+	+

Инструменты Process Mining

«Источники» данных о бизнес-процессах

Возможности Process Mining

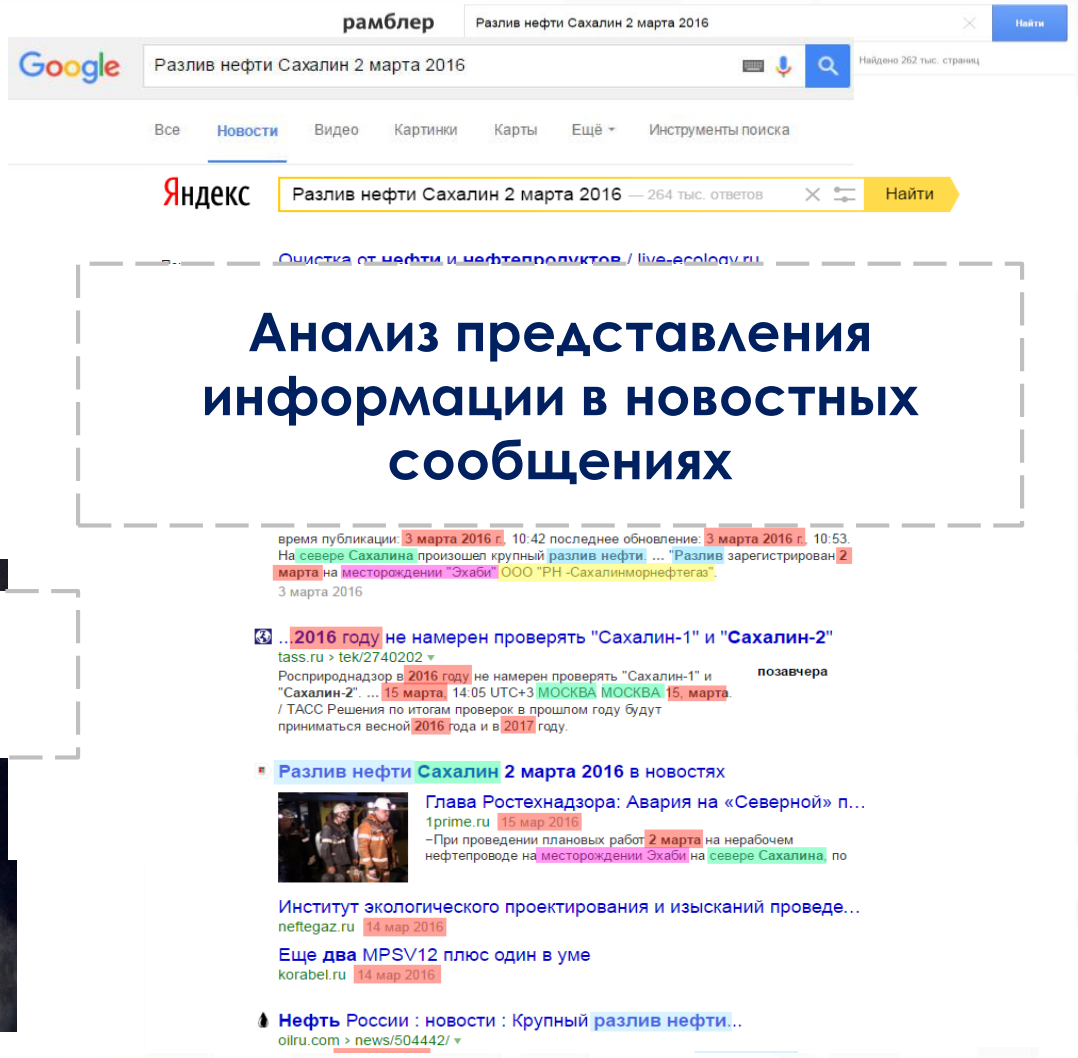
ProM поддерживает весь спектр современных задач Process Mining



Анализ предметной области

Формализация модели предметной области

Анализ представления информации в новостных сообщениях

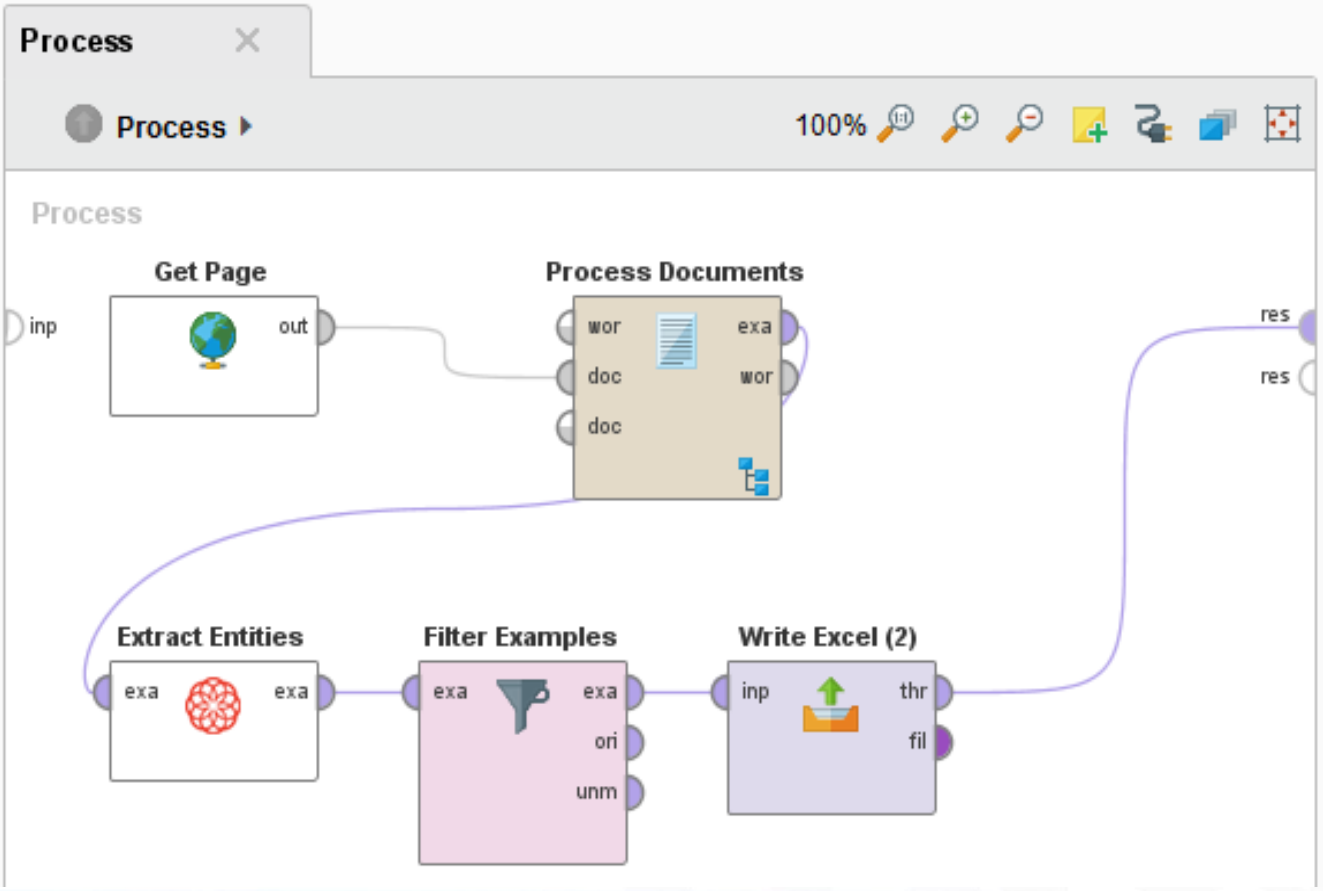


Основные типы событий и ключевые атрибуты

- ***Disaster*** *(date, oil company, place)* – сами катастрофы: пожар, разлив.
- ***Financial implication*** *(organization)* – оценки финансового ущерба, сюда относятся как затраты по устранению последствий, так и другие экономические показатели предприятий, населения, стран.
- ***Industry news*** *(oil company, publication date)* – возможные научные и технологические открытия, достижения, инновации, либо информация, связанная с функционированием компаний: расширение, банкротство…
- ***Sanction*** *(date)* – информация о санкциях, штрафах.
- ***Socio-environmental implication*** *(publication date)* – влияние на население, жертвы, ущерб сельскому хозяйству, влияние на общество.
- ***Socio-political*** *(date, place)* – влияние на политику государства, изменение взаимоотношений, влияние на внешнюю торговлю, морские пути.
- ***Noise*** – шумы.

Извлечение и обработка данных

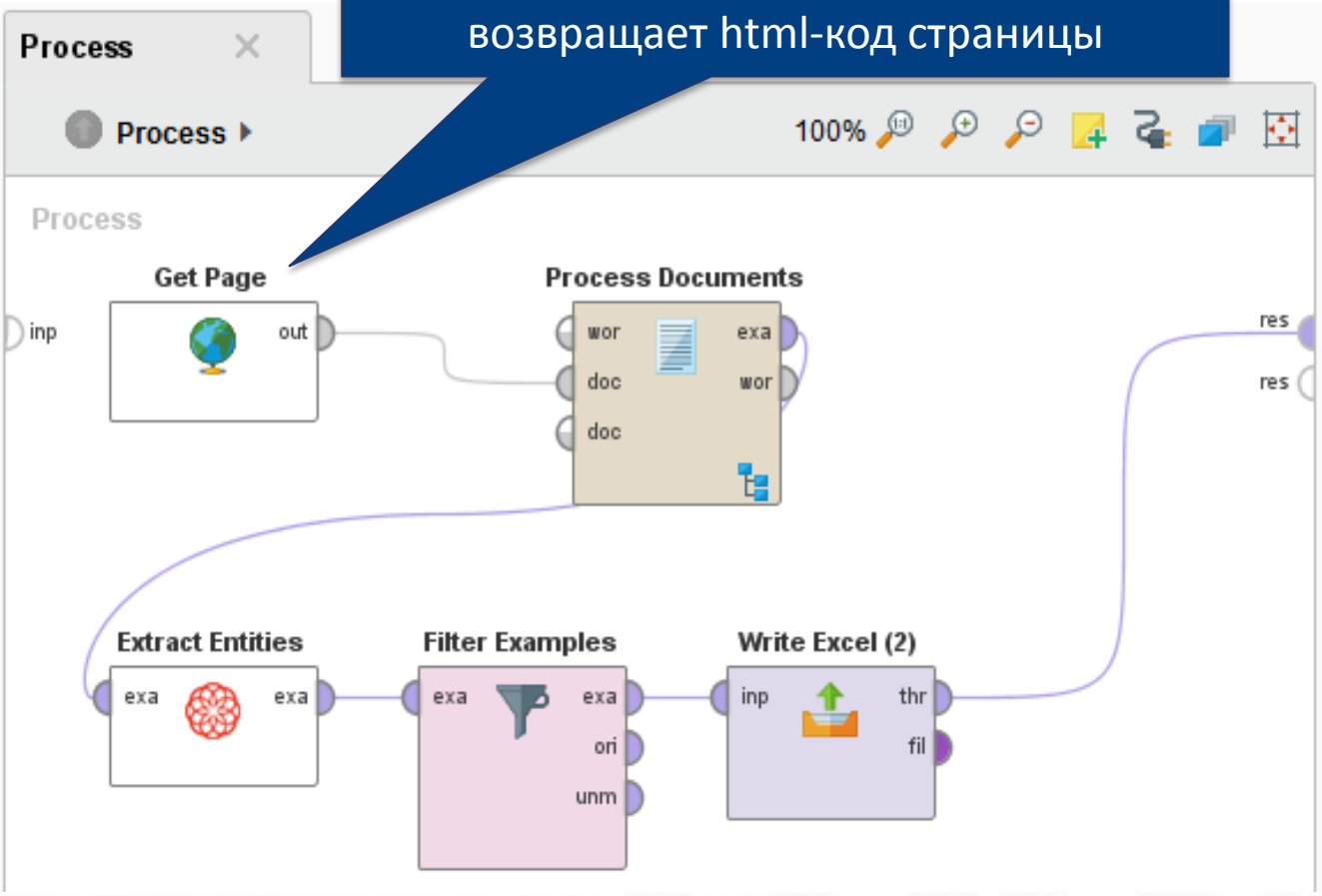
rapidminer



Извлечение и обработка данных

rapidminer

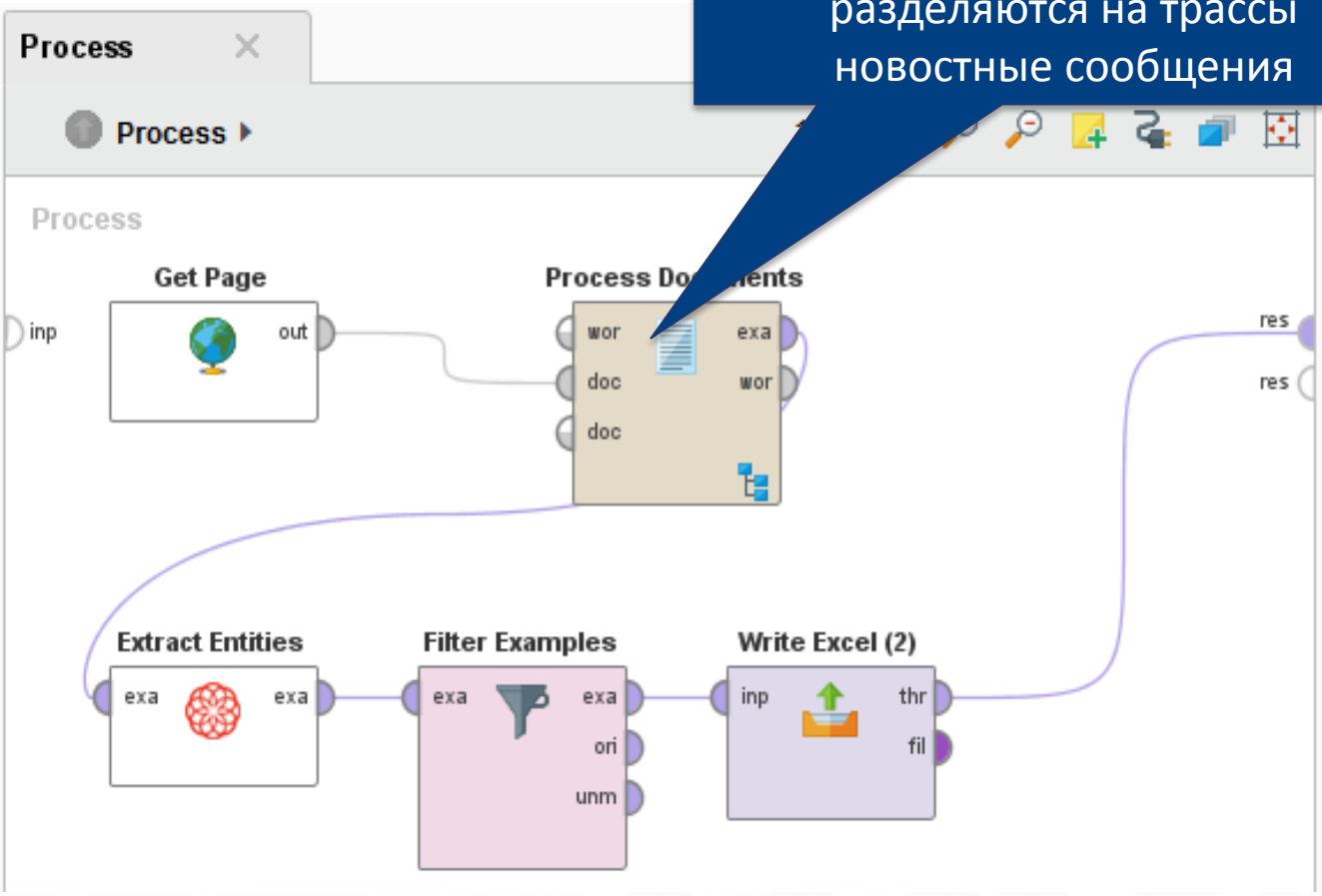
Оператор Get Page принимает на вход ссылку на Интернет-источник и возвращает html-код страницы



Извлечение и обработка данных

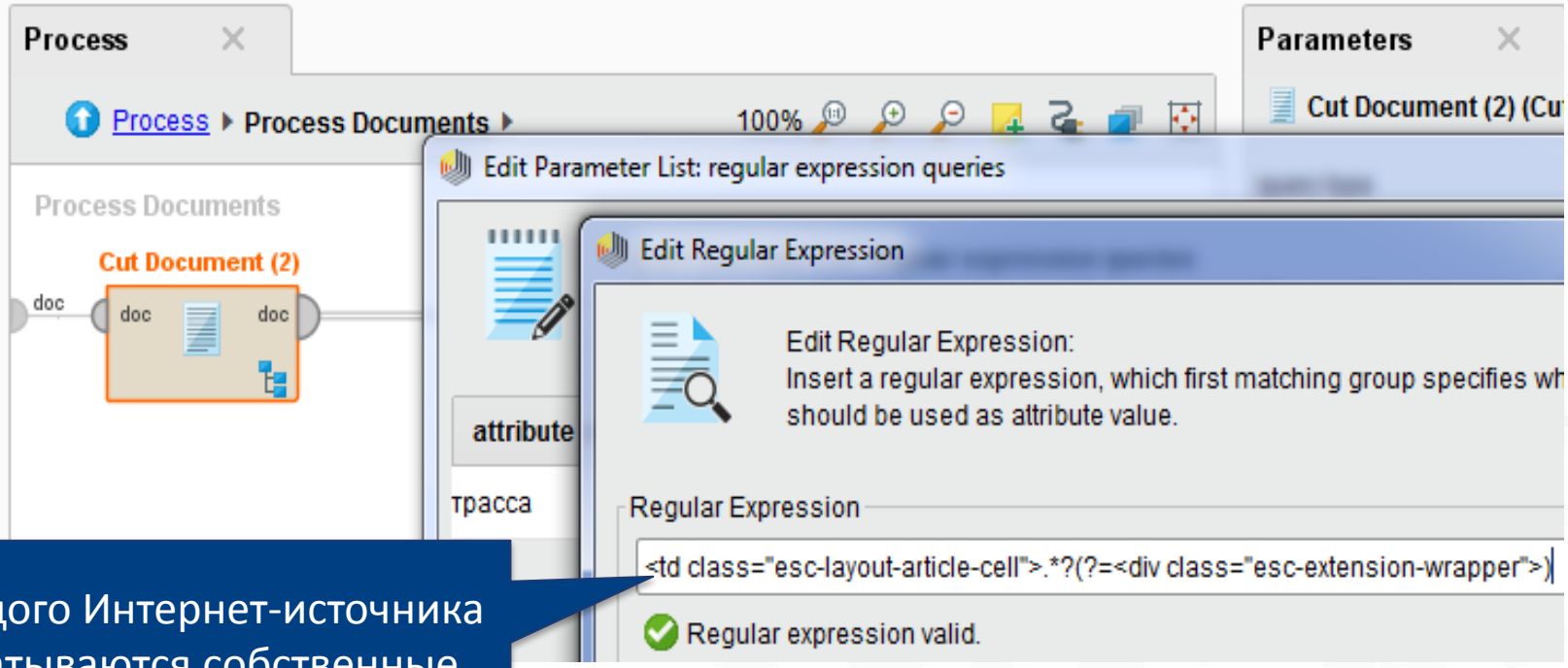
rapidminer

С помощью регулярных выражений вычленяются и разделяются на трассы новостные сообщения



Извлечение и обработка данных:
обработка документа, выделение трассы процесса

rapidminer



Для каждого Интернет-источника разрабатываются собственные регулярные выражения

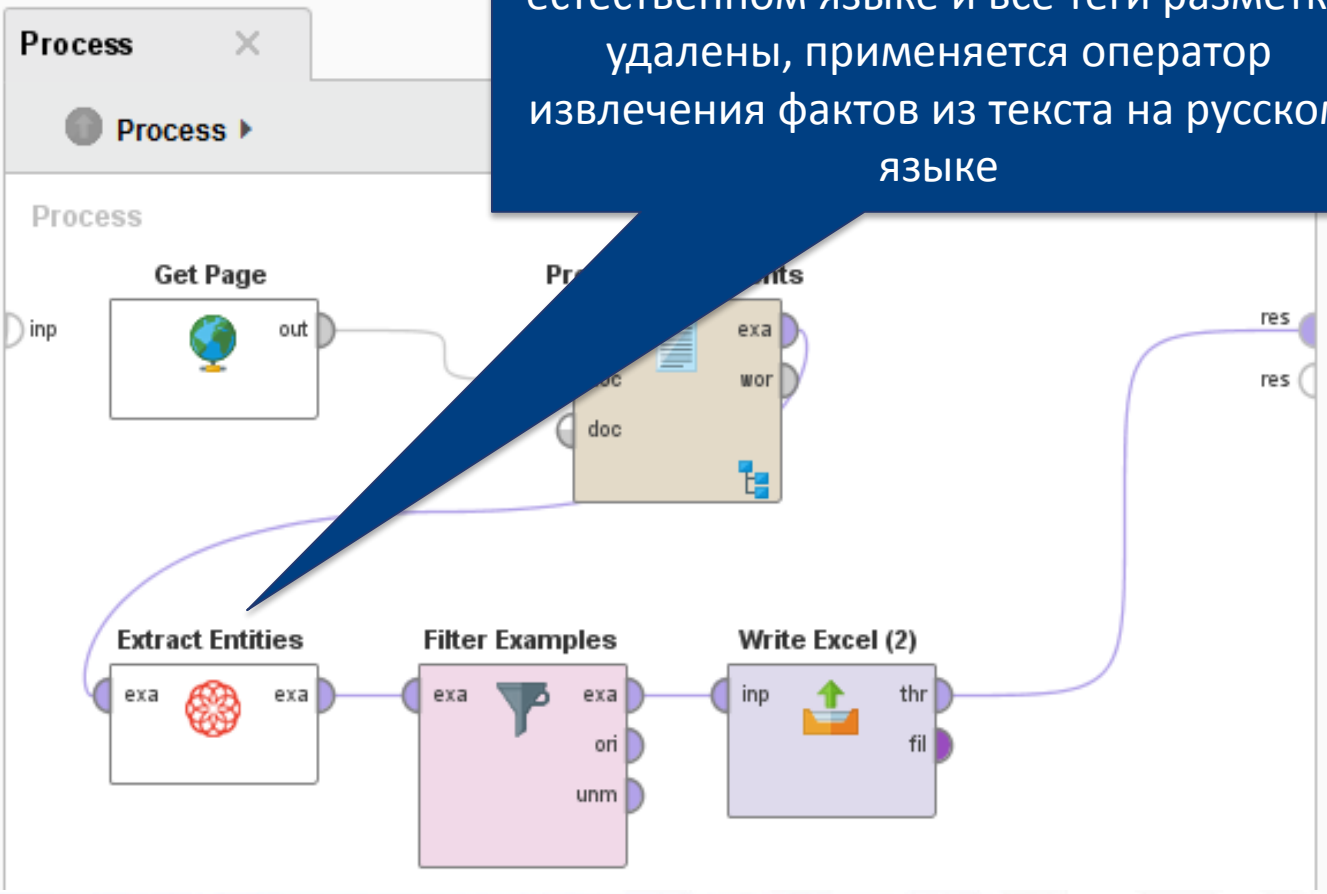
Извлечение и обработка данных

Row No.	Новости	URL	Content-Type	query_key	Источник	Заголовок	Дата публикации	Текст
1	https://news....	https://news....	text/html; char...	трасса	Investing.com Россия	Минприроды будет...	13 мая 2016 г.	Минприроды надее...
2	https://news....	https://news....	text/html; char...	трасса	Пронедра	МИД Украины зая...	12 мая 2016 г.	Соответствующая н...
3	https://news....	https://news....	text/html; char...	трасса	Няръяна вындер	<b>Разлив нефти<...	4 часа назад	<b>Разлив нефти</...
4	https://news....	https://news....	text/html; char...	трасса	Новости природы	<b>Разлив нефти<...	10 мая 2016 г.	В связи с загрязнен...
5	https://news....	https://news....	text/html; char...	трасса	Нефть России	Экологический сов...	11 мая 2016 г.	Федеральный эксп...
6	https://news....	https://news....	text/html; char...	трасса	Росбалт.RU	Коренные жители ...	6 мая 2016 г.	Как сообщает журн...
7	https://news....	https://news....	text/html; char...	трасса	Росбалт.RU	Сразу в двух район...	5 мая 2016 г.	В Курманаевском и ...
8	https://news....	https://news....	text/html; char...	трасса	Вектор News	За <b>разлив неф...	4 мая 2016 г.	В частности племен...
9	https://news....	https://news....	text/html; char...	трасса	НеваИнфо	В Мексиканском з...	12 мая 2016 г.	Как уточняется, <b>...
10	https://news....	https://news....	text/html; char...	трасса	В городе	Вблизи Севастопо...	11 мая 2016 г.	В акватории Севаст...
11	https://news....	https://news....	text/html; char...	трасса	UGRA-NEWS	В Сургуте на станц...	13 мая 2016 г.	На станции юных н...
12	https://news....	https://news....	text/html; char...	трасса	Oil.Эксперт	В Мексиканском з...	13 мая 2016 г.	Более 337,8 тыс. ку...

Извлечение и обработка данных

rapidminer

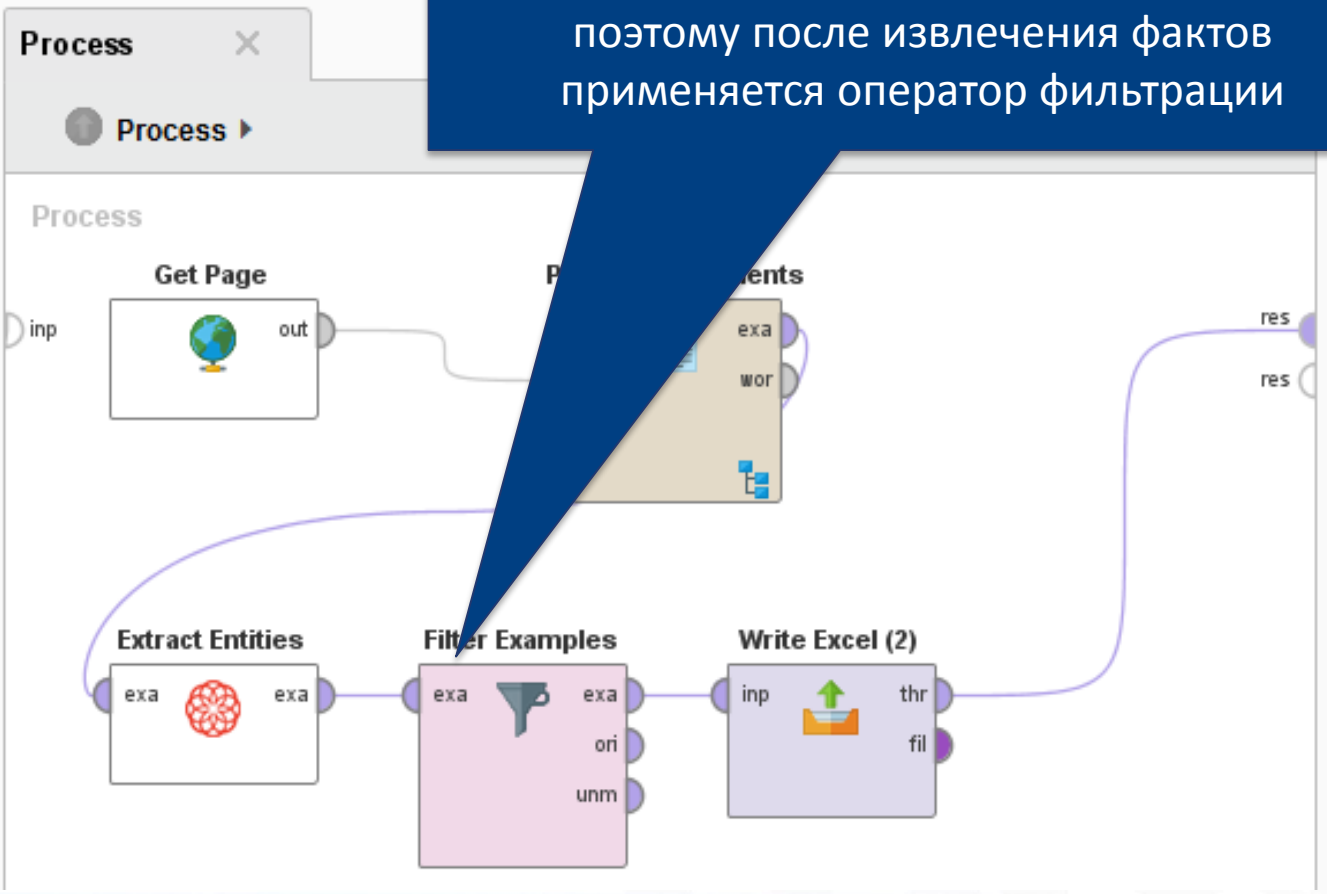
После того как получен текст на естественном языке и все теги разметки удалены, применяется оператор извлечения фактов из текста на русском языке



Извлечение и обработка данных

rapidminer

В системе не реализована возможность добавления пользовательских словарей, поэтому после извлечения фактов применяется оператор фильтрации



Извлечение и обработка данных:
полученные результаты

rapidminer

A	B	C	D	E
Источник	**Дата публикаци**	**Entity**	**EntityType**	**Entity(2)**
Slon.ru	26.05.2016	Shell	ORGANIZATION	подали в суд
Neftegaz.ru	26.05.2016	Shell	ORGANIZATION	завершают работы
Neftegaz.ru	26.05.2016	Shell	ORGANIZATION	ликвидации
ФБА "Экономика сегодня"	24.05.2016	Shell	ORGANIZATION	собираются обратиться в суд
ФБА "Экономика сегодня"	24.05.2016	Shell	ORGANIZATION	взыскать компенсацию
ТЭКНОБЛОК	24.05.2016	Shell	ORGANIZATION	Разлив проищошел
Вести Экономика	24.05.2016	Shell	ORGANIZATION	замечен разлив нефти
Вести Экономика	24.05.2016	Shell	ORG	
Вести Экономика	24.05.2016	Shell	OR	
Вести Экономика	24.05.2016	Shell	OR	

A	B	C	D
Источник	**Дата публикации**	**Entity**	**EntityType**
Росбалт	26.05.2016	РН-Юганскнефтегаз	ORGANIZATION
ЮграPRO	24.05.2016	Юганскнефтегаз	ORGANIZATION
Росбалт	26.05.2016	Лукойл	ORGANIZATION
ПитерБургер.ru	24.05.2016	Роснефть	ORGANIZATION
ПитерБургер.ru	24.05.2016	BP	ORGANIZATION
РИА Новости	11.04.2016	TransCanada Corp	ORGANIZATION
РИА Новости	04.03.2016	Роснефть	ORGANIZATION
National Geographic I	18.05.2016	Royal Dutch Shell	ORGANIZATION

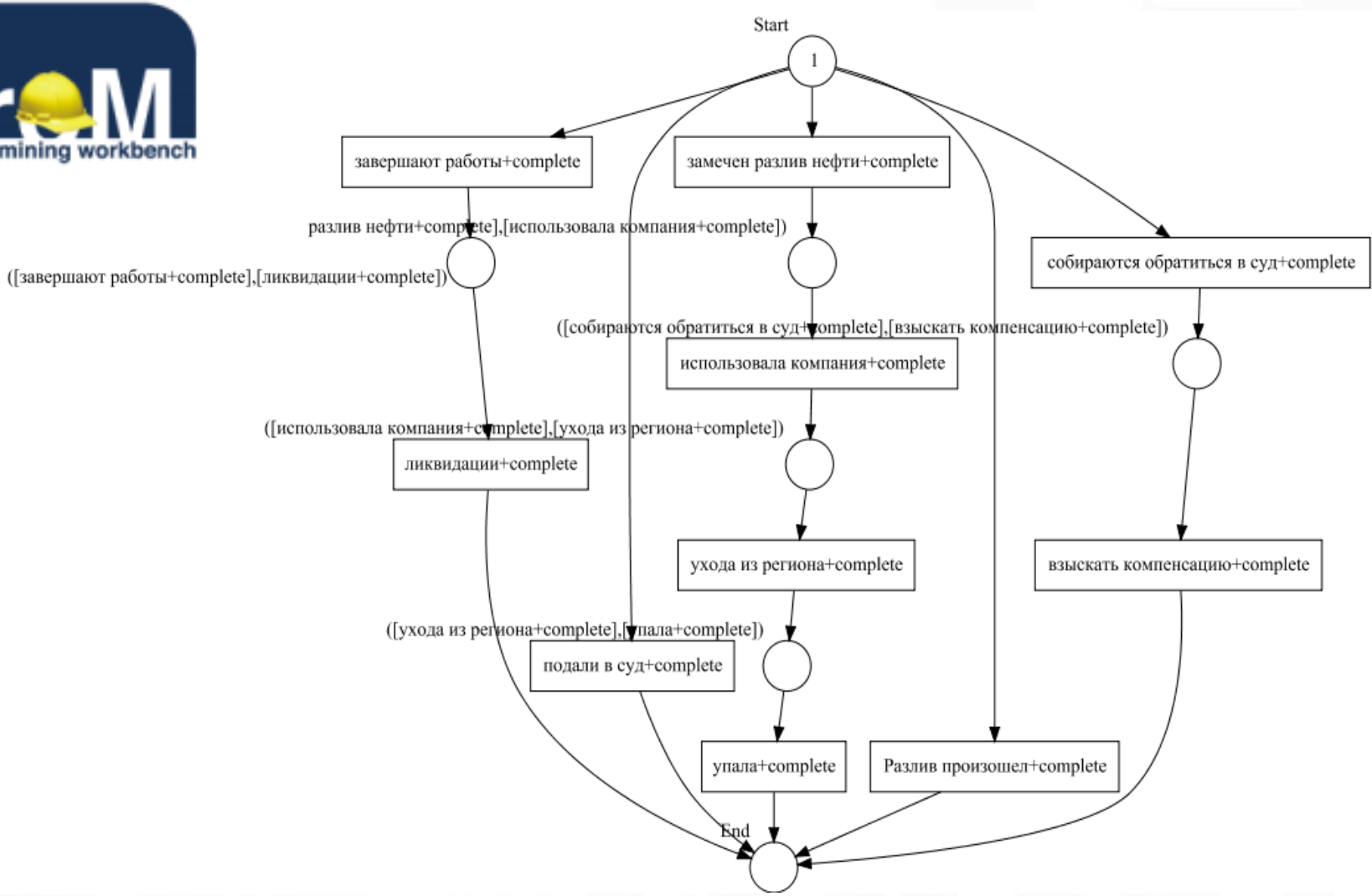
Извлечение и обработка данных:
полученные результаты

Извлеченные события и их атрибуты могут быть импортированы в ProM с помощью компонента XESame, для этого необходимо определить соответствие атрибутов в таблице и стандартных расширений лога.

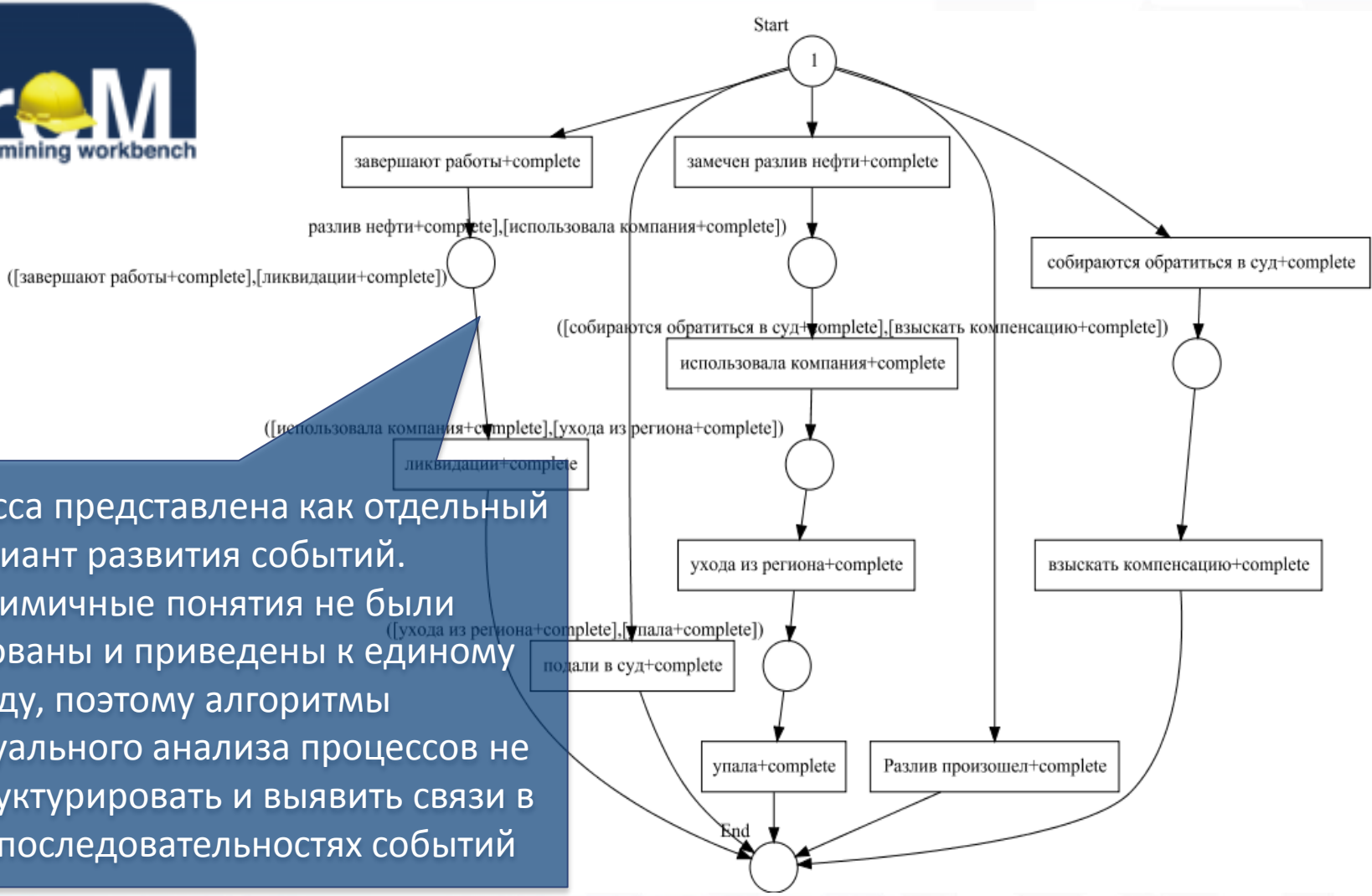
A	B	C	
Источник	**Дата публикаци**	**Entity**	**Ent**
Slon.ru	26.05.2016	Shell	OR
Neftegaz.ru	26.05.2016	Shell	ORGANIZATION
Neftegaz.ru	26.05.2016	Shell	ORGANIZATION
ФБА "Экономика сегодня"	24.05.2016	Shell	ORGANIZAT
ФБА "Экономика сегодня"	24.05.2016	Shell	ORGANIZ ON
ТЭКНОБЛОК	24.05.2016	Shell	ORGANIZATION
Вести Экономика	24.05.2016	Shell	ORGANIZATION
Вести Экономика	24.05.2016	Shell	ORGANIZATION
Вести Экономика	24.05.2016	Shell	OR
Вести Экономика	24.05.2016	Shell	OR

Источник	**Дата публикации**	**Entity**	**EntityType**
Росбалт	26.05.2016	РН-Юганскнефтегаз	ORGANIZATION
ЮграPRO	24.05.2016	Юганскнефтегаз	ORGANIZATION
Росбалт	26.05.2016	Лукойл	ORGANIZATION
ПитерБургер.ru	24.05.2016	Роснефть	ORGANIZATION
ПитерБургер.ru	24.05.2016	BP	ORGANIZATION
РИА Новости	11.04.2016	TransCanada Corp	ORGANIZATION
РИА Новости	04.03.2016	Роснефть	ORGANIZATION
National Geographic	18.05.2016	Royal Dutch Shell	ORGANIZATION

Построение формальной модели процесса в ProM



Построение формальной модели процесса в ProM



Каждая трасса представлена как отдельный вариант развития событий.
Синонимичные понятия не были сгруппированы и приведены к единому виду, поэтому алгоритмы интеллектуального анализа процессов не смогли структурировать и выявить связи в трассах и последовательностях событий

Результаты эксперимента

В ходе эксперимента выявлены ***следующие недостатки***:

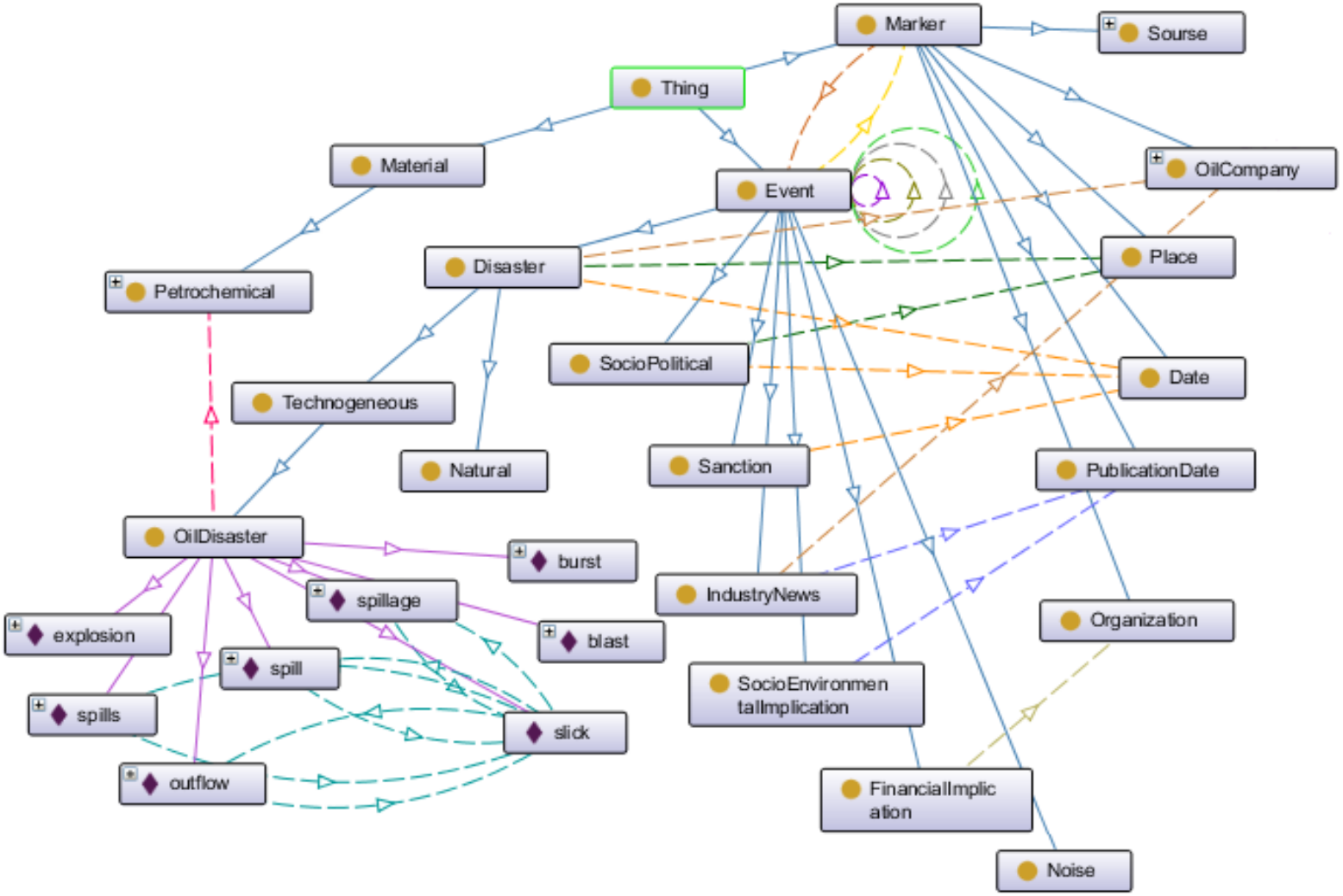
- Отсутствие возможности выявления причинно-следственных связей.
- Большое количество ошибок и потерь при извлечении событий.
- Отсутствие возможностей создания перечня атрибутов и их наборов без необходимости программной разработки дополнительных операторов.
- …

Развитие подхода

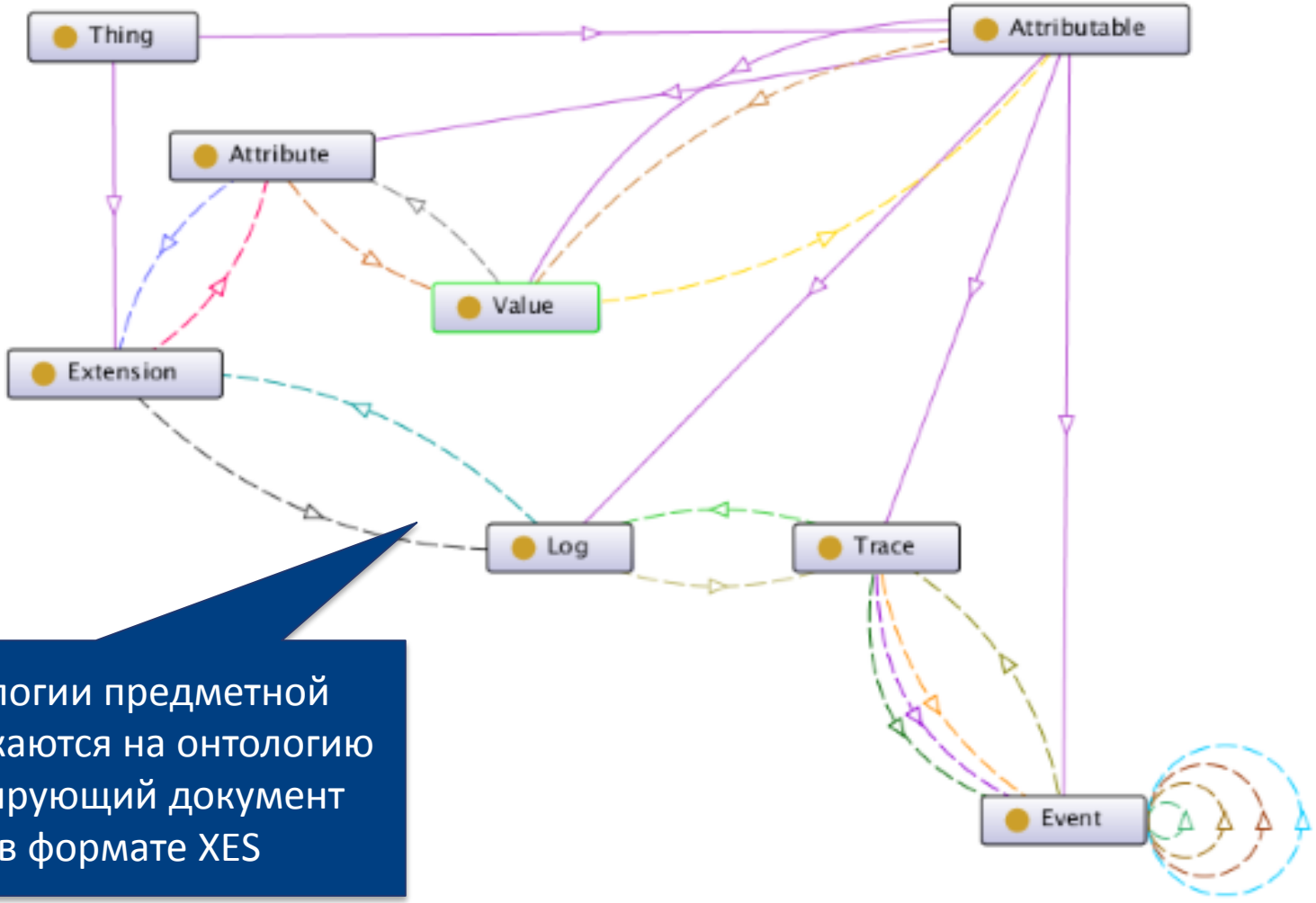
Для устранения недостатков предлагается:

- Использовать визуальный язык для разработки регулярных выражений.
- Применить средства синтаксического разбора текстов новостей на основе синтаксических шаблонов.
- Использовать онтологии для поиска данных и обработки логов.
- Реализовать алгоритм автоматического наполнения базы новостей на основе параметров, заданных пользователем.
- Для представления модели использовать DSL.
- …

Развитие подхода: онтология предметной области



Развитие подхода: онтология логов (фрагмент онтологии XES)



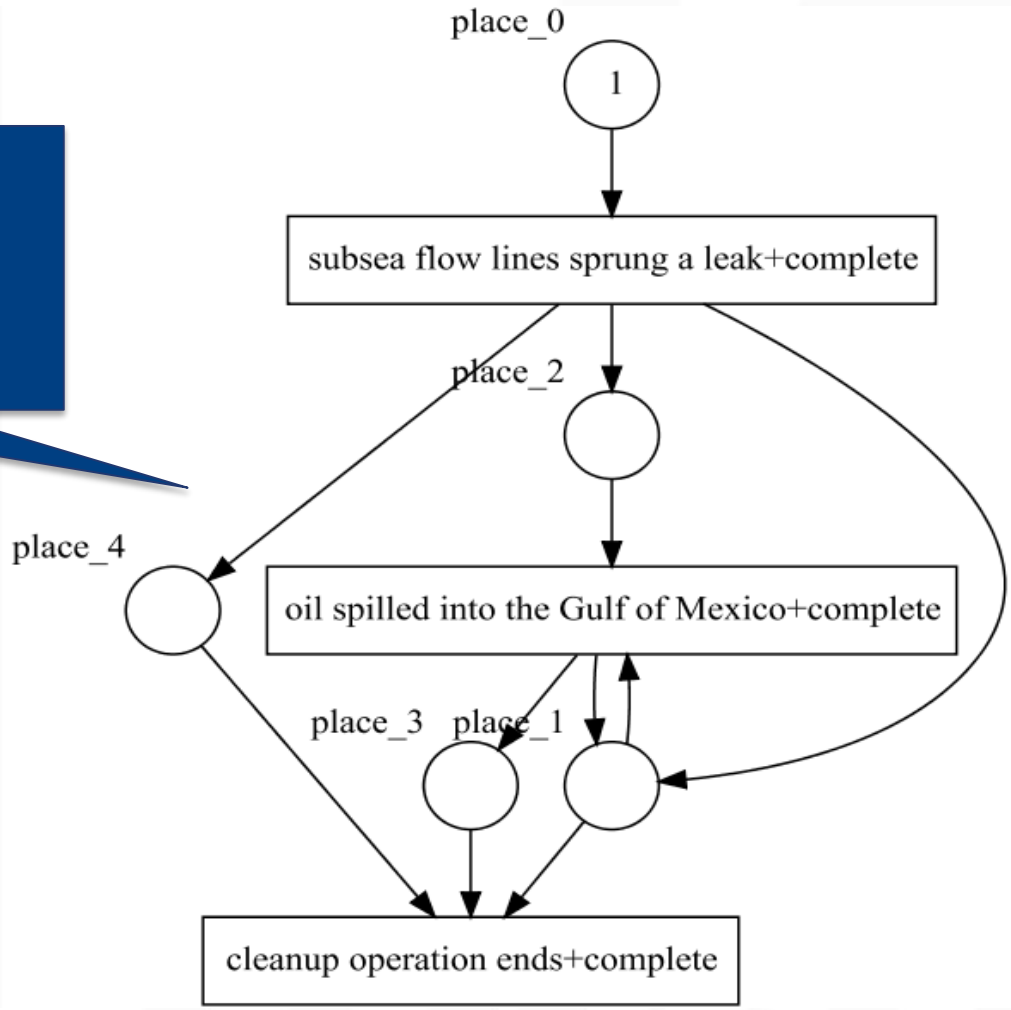
Понятия онтологии предметной области отображаются на онтологию логов. Результирующий документ создаётся в формате XES

Результирующий лог (фрагмент)

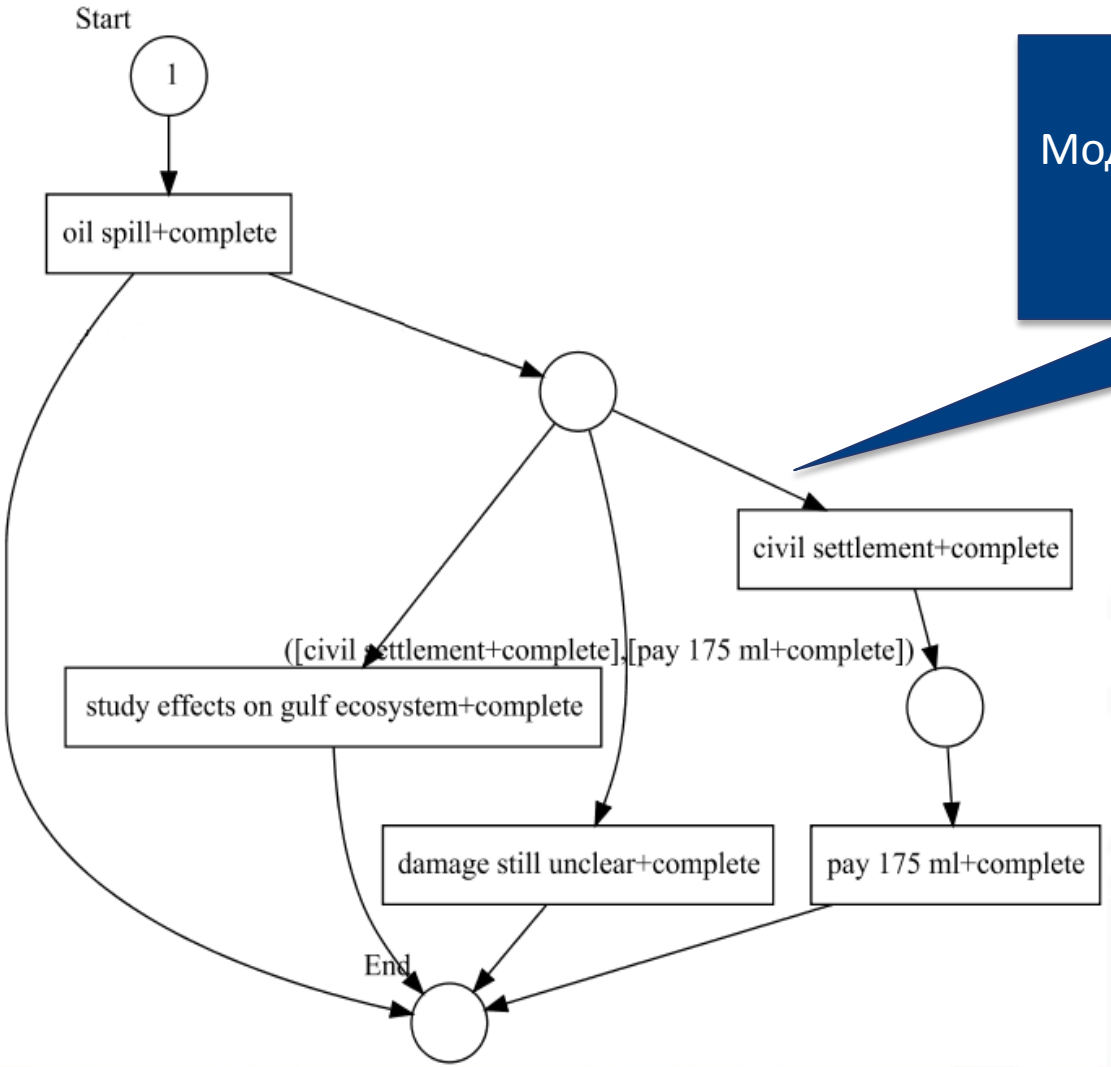
```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- This file has been generated with the OpenXES library. It conforms --
<!-- to the XML serialization of the XES standard for log storage and --
<!-- management. --
<!-- XES standard version: 1.0 --
<!-- OpenXES library version: 1.0RC7 --
<!-- OpenXES is available from http://www.openxes.org/
-->
<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="excercise1.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>
                .......
        </trace>
</log>
```

Модель процесса, построенная по уточнённому логу

Модель, построенная для запроса “Shell Spills Oil in the Gulf”



Модель процесса для запроса с изменёнными параметрами



Модель, построенная для запроса без указания места события

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Спасибо за внимание!