Pullenti 1.64. Именованные сущности.

# Обзор

Под именованной сущностью (ИС) будем понимать информационный объект с набором атрибутов, значения которых однозначно идентифицируют объект в рамках некоторого текста. Причём сам набор атрибутов определяется типом сущности (персона, организация …), а значения атрибутов отличают один объект от другого объекта этого же типа.

Вопрос отождествления сущностей (так называемая задача кореференции) является первостепенным: мало определить места в тексте, где встречаются сущности, важно правильно выделить и нормализовать их атрибуты, и, как следствие, правильно установить факт эквивалентности. Например, «… Иванова Ивана Ивановича … И.И.Иванов …».

Атрибуты у ИС бывают простые (строки, числа) и ссылки на другие ИС. Благодаря ссылкам множество выделенных ИС из текста в общем случае представляет собой ориентированный граф, в узлах которого находятся ИС.

В Pullenti ИС выделяются так называемыми анализаторами (подробности см. документацию для разработчиков). Есть анализаторы общего типа, выделяющие такие объекты, как персоны, организации, даты, географические объекты, денежные суммы и пр. Есть специализированные анализаторы, которые покрывают некоторую предметную область. Например, выделяющие структуру нормативно-правового акта и договора с его реквизитами, анализирующие титульные листы, литературных персонажей, происшествия и пр. Список анализаторов и выделяемых ими типов ИС пополняется исходя из потребностей реальных задач.

А что не является ИС?

Например, текст «Мама мыла раму» не содержит в общем случае ИС, и как персона МАМА не определяется ввиду отсутствия нужных атрибутов. Однако с точки зрения специализированного анализатора «Семантика» здесь будет 3 сущности (предикат «МЫТЬ» и объекты «МАМА» и «РАМА»), но здесь похожий результат будет и в случае «Тряпка мыла раму».

Текст «Маша мыла раму» также не содержит ИС, так как для выделения персон нужно больше информации. Например, хотя бы фамилия или ещё какой атрибут. Для текста при анализе на ИС составляется как бы микро-досье, и если персона имеет только имя, то этого мало и ИС не определяется. Иначе возникает много «шума», когда выделяются ИС там, где их реально нет.

Система при выделении ИС руководствуется базовым принципом: **приоритет качества перед полнотой**. То есть если есть сомнения в том, есть ли здесь ИС, то считаем, что её нет.

Итак, **понятие ИС относительное**, зависит от анализатора и его специализации. С точки зрения одного анализатора в конкретном месте может быть ИС, с точки зрения другого – нет.

# Анализаторы и список сущностей

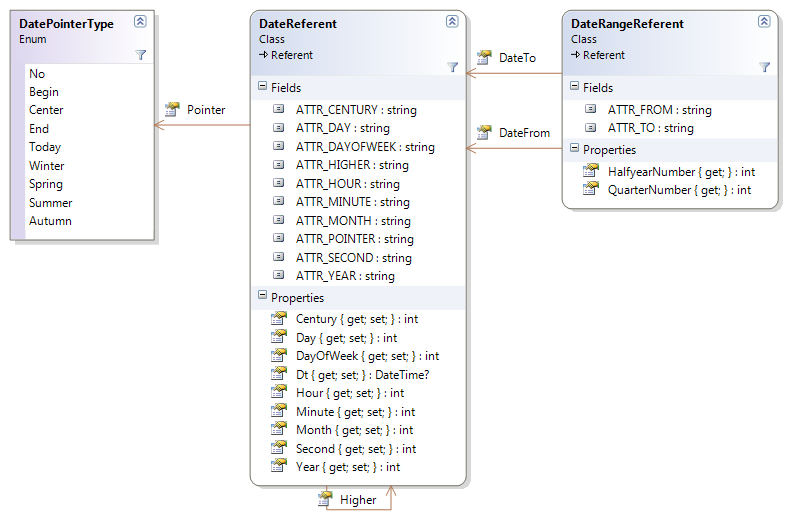
Приведём в таблице перечень ИС, выделяемых анализаторами на текущий момент. Названия анализаторов выводятся в одноклеточных строках. Для сущностей, отмеченных звёздочкой, реализация является предварительной (то есть не было реальных задач или разработка велась в рамках пилотных проектов, не вылившихся в релизы). В наименованиях указано значение, которое помещается в свойство TypeName у Referent (см. документацию разработчика).

Сущности класса «Общий» выделяются всегда из любого текста, класса «Специализированный» - только при специальном указании нужному анализатору (в дополнение к выделенным общим сущностям).

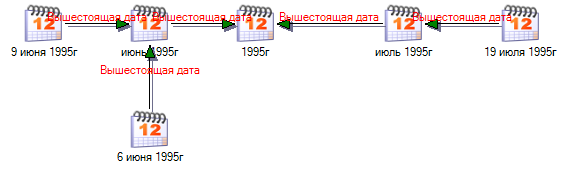
|  |  |  |
| --- | --- | --- |
| *Наименование ИС* | *Класс* | *Описание* |
| **EP.Analyzer.Date** | | |
| DATE | Общ. | Дата-время, есть поддержка относительных дат |
| DATERANGE | Общ. | Диапазон дат |
| **EP.Analyzer.Misc** | | |
| PHONE | Общ. | Телефоны |
| URI | Общ. | Сайты, электронные адреса и пр. (всё, что укладывается в структуру СХЕМА:ЗНАЧЕНИЕ – например, УДК, ИНН, ОКАТО…) |
| MONEY | Общ. | Денежные суммы |
| BANKDATA | Общ. | Банковские реквизиты |
| KEYWORD | Спец. | Ключевые комбинации (слова и словосочетания, ранжированные с точки зрения «важности» в тексте) |
| DEFINITION | Спец. | Определения (например, «под именованной сущностью понимается информационный объект …»), определяется сам термин и текст его определения. |
| DENOMINATION | Общ. | Специфические буквенно-числовые комбинации типа А-4, 1С, С300. |
| **EP.Analyzer.Location** | | |
| GEO | Общ. | Страны, регионы (области, районы) и населённые пункты. |
| STREET | Общ. | Элемент адреса (улица, площадь …) |
| ADDRESS | Общ. | Географический адрес |
| **EP.Analyzer.Fias** | | |
| FIAS | Спец. | Привязка адреса к объекту базы ГИС ФИАС |
| **EP.Analyzer.Organization** | | |
| ORGANIZATION | Общ. | Организации |
| **EP.Analyzer.Person** | | |
| PERSON | Общ. | Персона |
| PEERSONPROPERTY | Общ. | Свойство персоны (должность, звание, обращение и пр.) |
| PERSONIDENTITY | Общ. | Паспортные данные и др. документы |
| MAIL | Спец. | Блок письма (для анализа структуры текстов писем и выделения информации из подписей) |
| **EP.Analyzer.Transport** | | |
| TRANSPORT | Общ. | Транспортные средства, имеющие номера, наименования и\или модели. |
| **EP.Analyzer.Decree** | | |
| DECREE | Общ. | Ссылки на нормативно-правовые акты (НПА) |
| DECREEPART | Общ. | Ссылки на фрагменты НПА |
| DECREECHANGE | Общ. | Описание внесения изменения в НПА |
| INSTRUMENT  INSTRBLOCK  INSTRPARTICIPANT | Спец. | Восстановление структуры текста НПА или договора, определение участников. |
| **EP.Analyzer.Business** | | |
| *BUSINESSFACT (\*)* | Спец. | Бизнес-факты (покупка, продажа …) |
| FUNDS | Спец. | Акция (биржевая). |
| **EP.Analyzer.Fact** | | |
| *FACT (\*)* | Спец. | Факты |
| **EP.Analyzer.Biblio** | | |
| TITLEPAGE | Спец. | Информация с титульной страницы (диссертации, научной статьи, отчёта и пр.) |
| BOOKLINK  BOOKLINKREF | Общ. | Ссылка на внешний источник (из списка литературы, например). |
| *CHARACTER (\*)*  *CHARLINK (\*)* | Спец. | Литературный персонаж |
| **EP.Analyzer.Goods** | | |
| GOOD  GOODATTR | Спец. | Атрибуты товара из его краткого наименования |
| **EP.Analyzer.Spec** | | |
| *NAMEDENTITY (\*)* | Общ. | Разные мелочи типа морей, озёр, планет, гор, памятников, зданий и пр., имеющих имена. |
| **EP.Analyzer.Event** | | |
| *ACCIDENT (\*)*  *ARTEFACT (\*)*  *OCCURENCE (\*)*  *PARTICIPANT (\*)* | Спец. | Анализ сводки происшествий |
| **EP.Analyzer.Sentiment** | | |
| *SENTIMENT (\*)* | Спец. | Эмоциональный анализ |
| **EP.Analyzer.Semantic** | | |
| *OBJECT (\*)*  *PREDICATE (\*)*  *ACTANT (\*)* | Спец. | Семантический анализ (представление текста в виде семантического графа) |
|  |  |  |

# Даты

Даты и диапазоны дат выделяются анализатором EP.Analyzer.Date. В общем случае дата может содержать век, год, месяц, день месяца, день недели, час, минуту, секунду, нечёткий указатель («в начале сентября», времена года).



Рассмотрим текст «6 и 9 июня, 19 июля 1995 г.». Результирующий граф сущностей – дат:



Отметим, что хотя в тексте только три даты, но сущностей выделяется 6: дополнительно для года и для двух месяцев. То есть восстанавливается как бы иерархия, конечные узлы представляют собой конечные даты. Что касается атрибутов, то, например, «6 июня 1995 года» имеет следующее:

*Тип объекта: Дата*

*Свойства:*

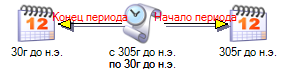
*День: 6*

*Связи с другими объектами*

*Вышестоящая дата: июнь 1995 года*

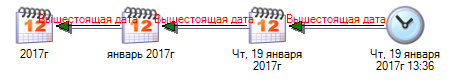
То есть ни месяц, ни год здесь явно не представлены, так как они подразумеваются в вышестоящей по иерархии сущности. Однако свойство Year или Month вернёт корректное значение, так как при отсутствии в атрибутах текущей сущности поиск производится в вышестоящей дате.

Рассмотрим текст «305 г. до н. э. - 30 г. до н. э.», здесь две даты и один диапазон:

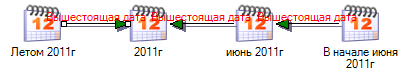


Отметим, что если год или век в атрибуте отрицательный, то считается до новой эры.

Вот так выглядит время в тексте «Четверг, 19 января 2017, 13:36 +03:00»:



Вот примеры нечётких указателей «начало июня 2011, летом 2011»:



Отметим, что анализатор EP.Analyzer.Date имеет дополнительную возможность (класс DateExToken) по выделению относительных дат и функцию приведения их к абсолютным датам относительно текущей даты. Например, «два дня назад», «в прошлом году», «за прошедшие 2 месяца», «первый квартал текущего года».