119270, Москва, Лужнецкая наб., д. 6,
стр.1, офис 214, ООО «ЭР СИ О»
Тел./факс: (495) 287-98-87
E-mail: info@rco.ru
http://www.rco.ru

Руководство администратора RCO

Настройка описаний объектов

Москва, 2007

В содержание данного документа могут быть внесены изменения без предварительного уведомления. Названия организаций, имена и даты, используемые в качестве примеров, являются вымышленными, если не оговорено обратное.

Содержание

Обзор

В этом руководстве содержатся основные сведения по настройке описаний именованных объектов (персон и организаций), используемых в программных продуктах **RCO**.

Наличие полного описания объекта в дополнение к общим лингвистическим правилам позволяет более точно идентифицировать в тексте все упоминания объекта, отождествляя различные способы его обозначения. Описания задаются в едином xml-формате и включаются в состав лингвистического обеспечения, используемого продуктами **RCO**, согласно соответствующему регламенту (см. документацию по конкретному продукту).

Общие правила

Описание объекта включает в себя набор строковых атрибутов, значения которых задаются пользователем.

Все объекты имеют атрибут, называемый *идентификатор объекта*, который содержит наименование объекта в произвольной, удобной для пользователя форме. Значение этого идентификатора используется в качестве обозначения объекта при выдаче результатов анализа текста в пользовательском интерфейсе.

Состав прочих атрибутов зависит от типа объекта («*персона*», «*организация*») и перечисляется далее.

При задании строковых значений атрибутов следует неукоснительно соблюдать ряд общих правил, как то:

1. Текстовая строка (слово или словосочетание) должна стоять в нормальной форме, т.е. в той форме, в которой принято писать заголовки словарных статей, за исключением атрибутов-синонимов, о чем сказано ниже. При этом первое существительное ставится в форму именительного падежа (обычно в единственном числе), а формы прилагательных согласуются с существительными (*Союз производителей черных и цветных металлов, металлургический комбинат, «Юкос», Иванов, Михаил, президент*). Указание формы множественного числа автоматически означает, что формы единственного числа при поиске в тексте следует игнорировать, например: «*Мобильные телесистемы*».

2. Текстовая строка – слово или словосочетание – должно быть написано с учетом следующих требований к регистру символов:

   a) имена нарицательные пишутся строчными буквами, за исключением прилагательных в названиях организаций, например, *Московский государственный университет*. Написание с заглавной буквы первого символа слова/словосочетания автоматически означает, что соответствующая строка в тексте обязана начинаться с заглавной буквы.

   b) Имена собственные пишутся с заглавной буквы (*Василий*, *Иванов*, «*Газпром*»). При этом неизменяемые названия организаций следует писать целиком заглавными буквами (*МДМ*, *НОСТА*). Слово или словосочетание, написанное латиницей, всегда означает неизменяемое название (*Microsoft*, *TWG*, *Raiffeisen Zentralbank Osterreich AG*).

В качестве синонимов к объектам не разрешается использовать имена объектов другого класса, например фамилия *Ходорковский* не может использоваться в качестве синонима к компании *«Юкос»* и наоборот.

Описание персоны

Описание объекта типа «*персона*» задается в xml-файле следующего формата:

```
<object id="ИдентификаторОбъекта" type="person">
<fields>
<field name="gender">Род</field>
<field name="last name" modify="Изменяемость">Фамилия</field>
<field name="first name" modify="Изменяемость">Имя</field>
<field name="middle name" modify="Изменяемость">Отчество</field>
</fields>
<desc>
<syn type="normal" case="Склонение">Синоним</syn>
<syn type="normal" case="Склонение">Синоним</syn>
...
<syn type="context" case="Склонение">КонтекстныйСиноним</syn>
<syn type="context" case="Склонение">КонтекстныйСиноним</syn>
...
<syn type="referent">РеферентныйКонтекст</syn>
<syn type="referent">РеферентныйКонтекст</syn>
...
</desc>
</object>
```

Описание объекта типа «*персона*» содержит атрибуты, перечисленные ниже. Любой из атрибутов, кроме идентификатора и рода, может отсутствовать. Некоторые атрибуты могут иметь много различных значений, что отмечено как *множ.*

Фамилия

Одно слово в нормальной форме, написанное с заглавной буквы. Слово имеет признак изменяемости.

Имя

Одно слово в нормальной форме, написанное с заглавной буквы. Слово имеет признак изменяемости.

Отчество

Одно слово в нормальной форме, написанное с заглавной буквы. Слово имеет признак изменяемости. Указывает, склоняется слово (имя, фамилия, отчество) или нет. Возможные значения: *yes, no*.

Род

Род персоны. Возможные значения: *мужской, женский*.

Синоним (множ.)

Возможное синонимичное обозначение объекта в тексте – слово или словосочетание, заданное в нормальной форме. Для персоны это обычно должность, прозвище. Например, синонимами к *Владимиру Путину* являются слова: *президент России, российский президент, российский лидер, кремлевская власть, Кремль*. При нахождении в тексте такого обозначения оно считается синонимом объекта, если в тексте нет информации о том, что это обозначение другого объекта. Так, *президент России* будет считаться обозначением *Владимира Путина*, если в тексте не сказано: *президент России Борис Ельцин, новый президент России* и т.п.

Контекстный синоним (множ.)

Возможное синонимичное обозначение объекта в тексте – слово или словосочетание, заданное в нормальной форме. В отличие от обычного синонима, контекстный синоним считается обозначением соответствующего объекта лишь в том случае, если ранее в тексте явно упоминался сам объект (по ФИО или синониму). Например, для *Аллы Пугачевой* контекстными синонимами могут быть слова: *певица*, *примадонна*, *звезда эстрады, звезда российской эстрады*. Так же, как и обычные синонимы, контекстные синонимы проверяются на отсутствие противоречий в тексте: *звезда эстрады* будет считаться обозначением *Аллы Пугачевой*, если в тексте не сказано: *звезда эстрады София Ротару, известная эстрадная певица*.

Референтный контекст (множ.)

Слова и словосочетания, заданные в нормальной форме, присутствие которых в тексте позволяет снять неоднозначность – распознать соответствующий объект в случае, если он обозначен только именем и отчеством, или же его наименование (ФИО, синоним) может относиться к другим объектам. Например, присутствие в тексте таких слов, как *президентская гонка*, *российский лидер* позволяет отождествить неизвестный объект, названный *Владимир Владимирович*, с *Владимиром Путиным*, если в тексте не сказано *Владимир Владимирович Маяковский* или нечто подобное. А в случае, если используются описания двух объектов – *Путина* и *Ельцина*, то присутствие этих слов позволяет отождествить неуказанного в тексте *российского президента* с *Путиным*, а присутствие слов *расстрел Белого дома* – с *Ельциным*. При задании референтных контекстов нежелательно использовать имена собственные, так как это может помешать нормальной обработке объектов, имеющих эти имена.

Тип склонения

Определяет, каким образом строятся все грамматические формы, соответствующие заданному синониму. Возможны следующие значения:

- «Нормальная форма» или отсутствие значения – в этом случае грамматические формы синонима во всех падежах будут построены программой автоматически по заданной нормальной форме, на основании словаря и правил грамматики русского языка. При этом предполагается, что в составе словосочетаний склоняется только первое имя существительное вместе с предшествующими ему прилагательными, а следующие за ним слова должны стоять в той форме, в которой указаны;
- «Неизменяемое» – грамматические формы синонима во всех падежах совпадают с указанной формой;
- «Именительный», «Родительный», … или «Предложный» – указанная форма синонима соответствует заданному падежу.

Возможность явного указания склонения синонима позволяет точно распознавать и правильно связывать с другими словами сколь угодно сложные наименования персоны, в том числе не укладывающиеся в структуру ФИО, например *Патриарх Московский и всея Руси Алексий II, Гассан Абдурахман ибн Хоттаб IV*. В этом случае в описании объекта необходимо оставить пустыми значения атрибутов *Фамилия*, *Имя*, *Отчество*, добавив в синонимы все варианты написания наименования с указанием каждого из шести русских падежей.

Описание организации

Описание объекта типа «*организация*» задается в xml-файле следующего формата:

```
<object id="ИдентификаторОбъекта" type="organization">
<fields>
<field name="gender">Род</field>
<field name="full name" modify="Изменяемость">ПолноеНазвание</field>
</fields>
<desc>
<syn type="normal" case="Склонение">Синоним</syn>
<syn type="normal" case="Склонение">Синоним</syn>
<...>
</desc>
</object>
```

Описание объекта типа «*организация*» содержит атрибуты, перечисленные ниже. Любой из атрибутов, кроме идентификатора, рода и полного названия, может отсутствовать. Некоторые атрибуты могут иметь много различных значений, что отмечено как *множ.*

Полное название

Полное название организации (*«Челябинский металлургический комбинат», компания «Мобильные телесистемы», нефтяная компания «Лукойл», «Национальный резервный банк», холдинговая компания «Интеррос», министерство транспорта РФ, корпорация Trans World Group*). Как видно из приведенных примеров, желательно отразить в полном названии род деятельности организации, используя слова типа *«завод», «банк», «компания»*, что будет задействовано при отождествлении с именем собственным косвенных обозначений организации в тексте. Нежелательно использовать в названии обозначения организационно-правовой формы: ООО, открытое акционерное общество и т.п.

Признак изменяемости

Указывает, склоняется слово или нет. Относится только к тому слову в названии, которое представляет имя собственное (*МТС* – неизменяемое, «*Интеррос*» – изменяемое). Возможные значения: *yes, no*.

Род

Грамматический род относится только к тому слову в названии, которое представляет имя собственное (*«Татнефть»* – женского рода, «*Интеррос*» – мужского рода). Возможны значения: *мужской, женский* или *средний*.

Синоним (множ.)

Возможное синонимичное обозначение объекта в тексте – слово или словосочетание, заданное в нормальной форме. Синоним должен иметь соответствующий объекту лексико-семантический разряд. Например: *МДМ, группа МДМ, группа МДМ, группа «Московский деловой мир», «Московский деловой мир», МДМ-банк, филиал МДМ-банка*.

Тип склонения

Определяет, каким образом строятся все грамматические формы, соответствующие заданному синониму. Возможны следующие значения:

- «Нормальная форма» или отсутствие значения – в этом случае грамматические формы синонима во всех падежах будут построены программой автоматически по заданной нормальной форме, на основании словаря и правил грамматики русского языка. При этом предполагается, что в составе словосочетаний склоняется только первое имя существительное вместе с предшествующими ему прилагательными, а следующие за ним слова должны стоять в той форме, в которой указаны;
- «Неизменяемое» – грамматические формы синонима во всех падежах совпадают с указанной формой;
- «Именительный», «Родительный», … или «Предложный» – указанная форма синонима соответствует заданному падежу.

Возможность явного указания склонения синонима позволяет точно распознавать и правильно связывать с другими словами в тексте сколь угодно сложные наименования организаций, включающие в себя слова любых частей речи и знаки препинания (*«Идущие вместе»*, *«Отечество – вся Россия»*).