ИСИДА-Т

Интеллектуальная система извлечения и анализа данных из текстов

Извлечение информации

Цель:

- извлечь значимую информацию определенного типа из (больших массивов) текста для дальнейшей аналитической обработки

Результат:

- структурированные данные (объекты+отношения)

Примеры предметных областей

- **Спортивные события:** *<победитель>, <проигравший>, <счет>, <место встречи>, <дата>…*
- **База данных о рынке жилья:** *<район>,<цена>, <количество комнат>, <контактный телефон>…*
- **Выпуск новых товаров:** *<производитель>, <дата выпуска>, <название товара> …*

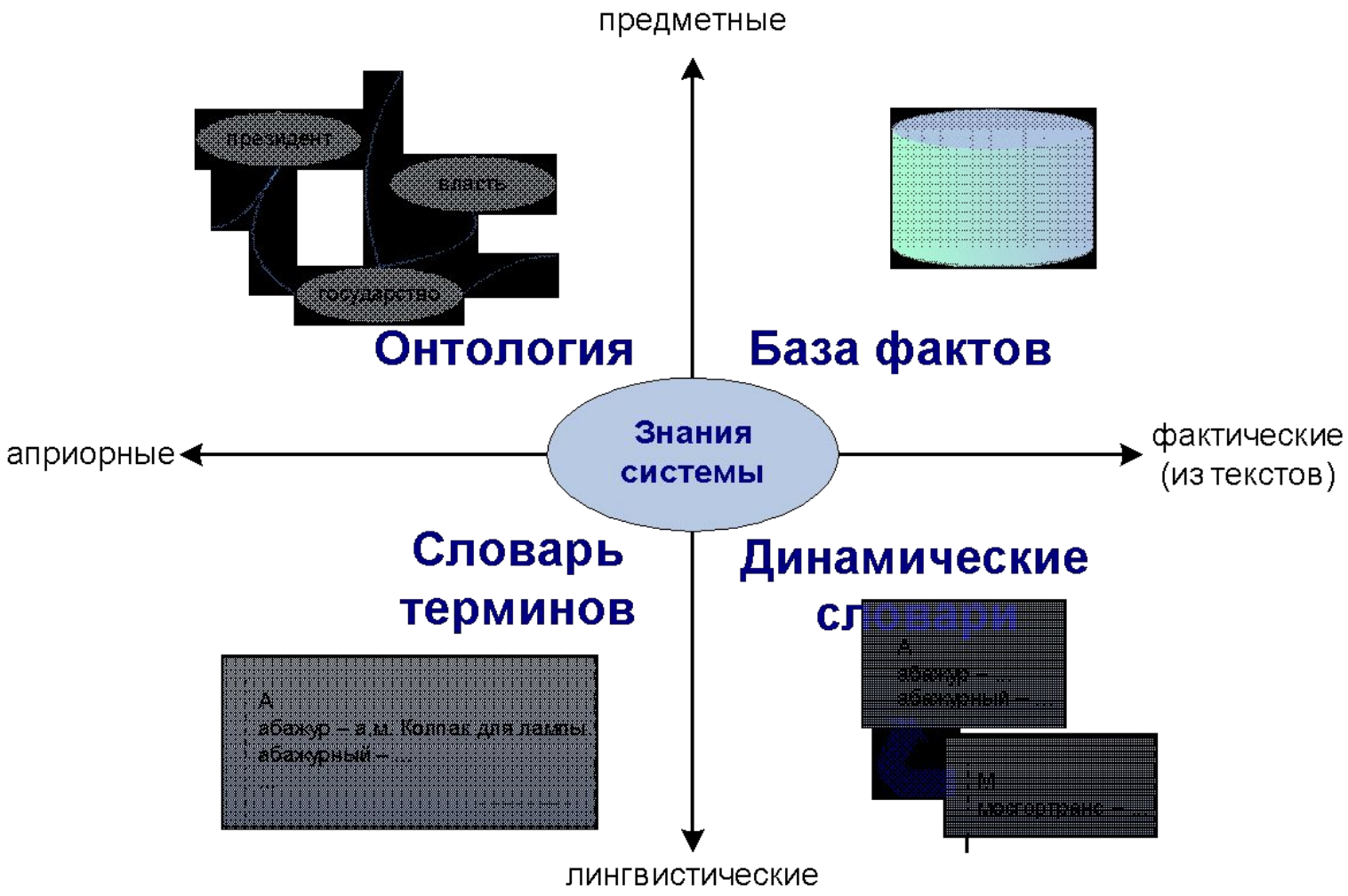
Приложения технологии извлечения информации

- семантическая кластеризация и классификация
- автоматическое аннотирование
- визуализация данных
- семантическое сравнение и поиск
- создание баз данных
- …

Основные компоненты системы

- Инфраструктурные службы (конфигурирование, параллельная обработка, взаимодействие модулей)
- Лингвистический процессор
- Интерпретатор правил извлечения информации
- Модули работы со знаниями предметной области

Знания в системе



Извлечение информации

- **В «слабом» смысле**
  - Обнаружение и пометка текстовых элементов и отношений (разметка текста)
- **В «сильном» смысле**
  - Переход от текстовых структур к модели предметной области

Извлечение информации в «слабом» смысле

- Лингвистическая обработка
  - Токенизация
  - Разбиение на предложения
  - Морфология
  - Частичный синтаксический анализ
- Словарное распознавание
- Распознавание именованных сущностей
- Частичный семантический анализ (в том числе, с использованием контекстных правил)

Построение первичных текстовых объектов

Примеры текстовых объектов

Тип:	**@лицо**	
Подтип:	**—**	
Атрибут	**Значение**	**Текстовый элемент**
Фамилия		*Ковальчук*
Имя		*Михаил*
Отчество		*Федорович*

Тип:	**@организация**	
Подтип:	**?**	
Атрибут	**Значение**	**Текстовый элемент**
Название		*Трансэк*
Правовая форма	AV@зао	*АОЗТ*

Тип:	**@организация**	
Подтип:	**@транспортная _компания**	
Атрибут	**Значение**	**Текстовый элемент**
Название		*Трансэк*
Правовая форма	AV@зао	*ЗАО*

Тип:	**@лицо**	
Подтип:	**—**	
Атрибут	**Значение**	**Текстовый элемент**
Фамилия		*Ковальчук*
Имя		*Михаил*

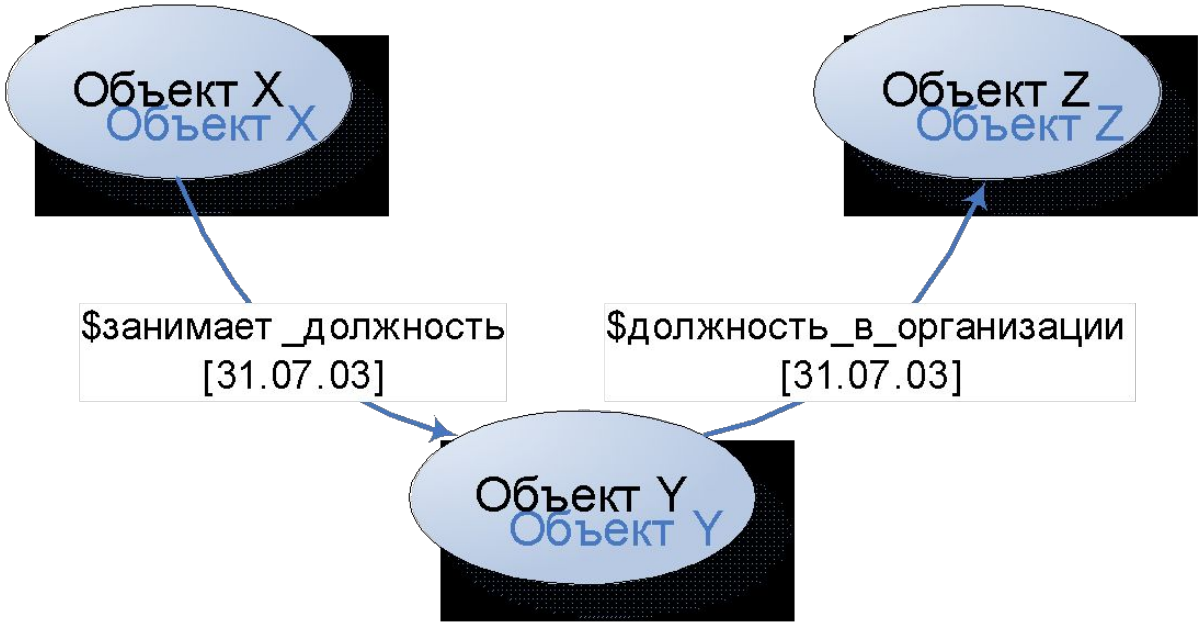
Тип:	**@должность**
Подтип:	**@генеральный _директор**

. . .

Всего по приведенным фрагментам построено 19 таких объектов.
Очевидно, что число **реальных объектов**, упомянутых в текстах, меньше.
Системе предстоит установить, какие из **текстовых объектов** соответствуют одному и тому же **реальному объекту**.

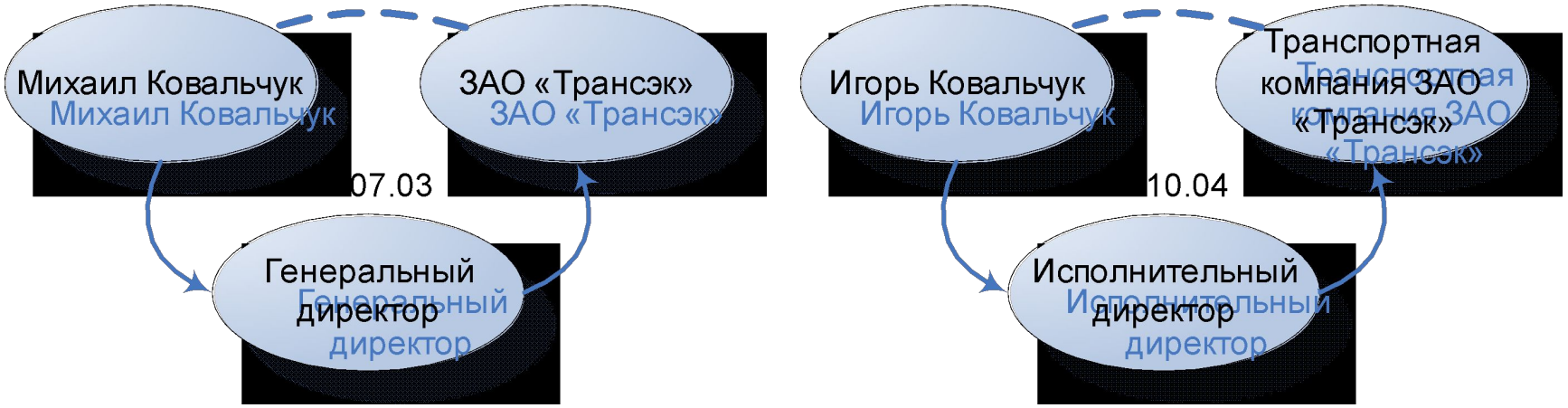
Построение текстовых фактов

- Текстовый факт — ситуация заданной структуры, имеющая временн'ую координату




```
($занимает_должность (Объект X, Объект Y), время: 31.07.03) &
($должность_в_организации (Объект Y, Объект Z), время: 31.07.03)
```

Построение текстовых фактов



Примеры построенных фактов

- *Михаил Ковальчук — генеральный директор ЗАО "Трансэк" [31.07.03]*
- *Михаил Ковальчук — член-корреспондент РАН, директор Российского научного центра "Курчатовский институт" [03.02.05]*
- *Игорь Ковальчук — исполнительный директор транспортной компании ЗАО «Трансэк» [06.10.04]*
- *Игорь Ковальчук — генеральный директор транспортной компании "Трансэк" [24.04.06]*

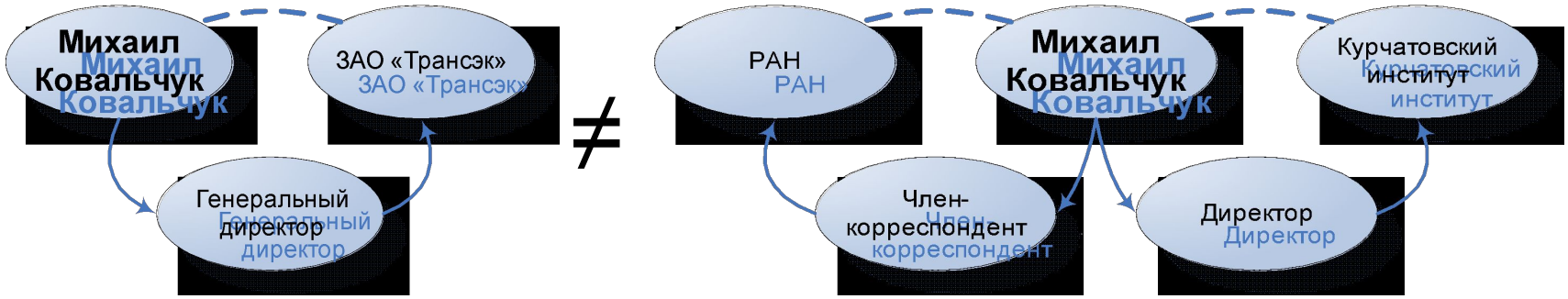
Установление кореферентности (примеры)

- Модуль собирает в один объект разбросанную по разным текстам информацию об организации «Трансэк»:

Тип:	**организация**
Подтип:	**транспортная _компания**
Название	Трансэк
Правовая форма	ЗАО
Профиль	автотранспортные услуги
Локализация	Санкт-Петербург

❗ Ни один отдельно взятый текст не содержал полного набора сведений об этой компании

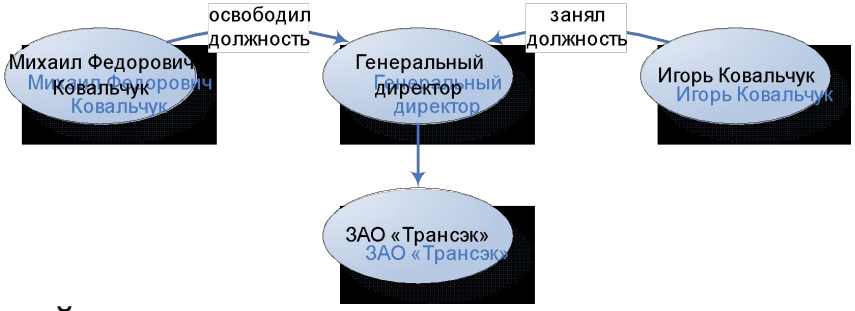
- Устанавливается, что существуют два разных лица с именем *Михаил Ковальчук*:



Вывод новых фактов

Пример вывода новых фактов об отставках и назначениях на основе данных, содержащихся в разных текстах

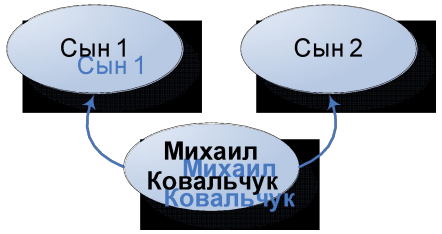
- «Смена лиц, занимающих должность»



- «Смена должностей лица»

Построение гипотез об отношениях между объектами из базы фактов

- По первому тексту система получает достоверный факт:



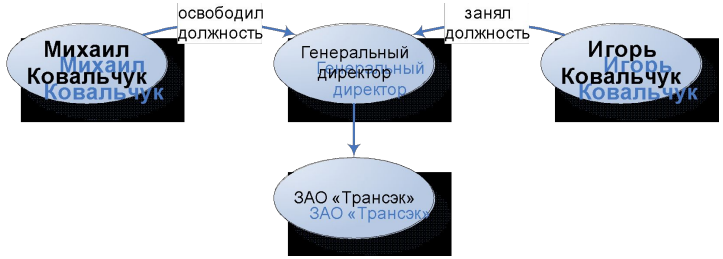
- Постулируется существование гипотетических объектов *Сын 1* и *Сын 2*, обладающих определенными свойствами, хоть и с разной степенью достоверности

- Поиск гипотетических объектов с такими свойствами в базе текстовых фактов обнаруживает два объекта:

- …

Построение гипотез… (окончание)

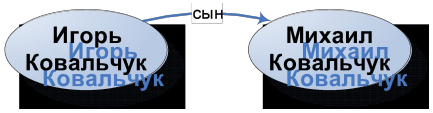
- Система ранее вывела факт:



- Предположим, в знаниях системы о мире есть фрагмент, который позволяет строить гипотезы — например, такого рода:

лицо, сменяющее однофамильца на руководящей должности с некоторой вероятностью состоит с ним в родственных отношениях

- Строится гипотеза:



- Для подтверждения или опровержения этой гипотезы у системы пока нет данных. Но они могут появиться по мере поступления новых текстов.

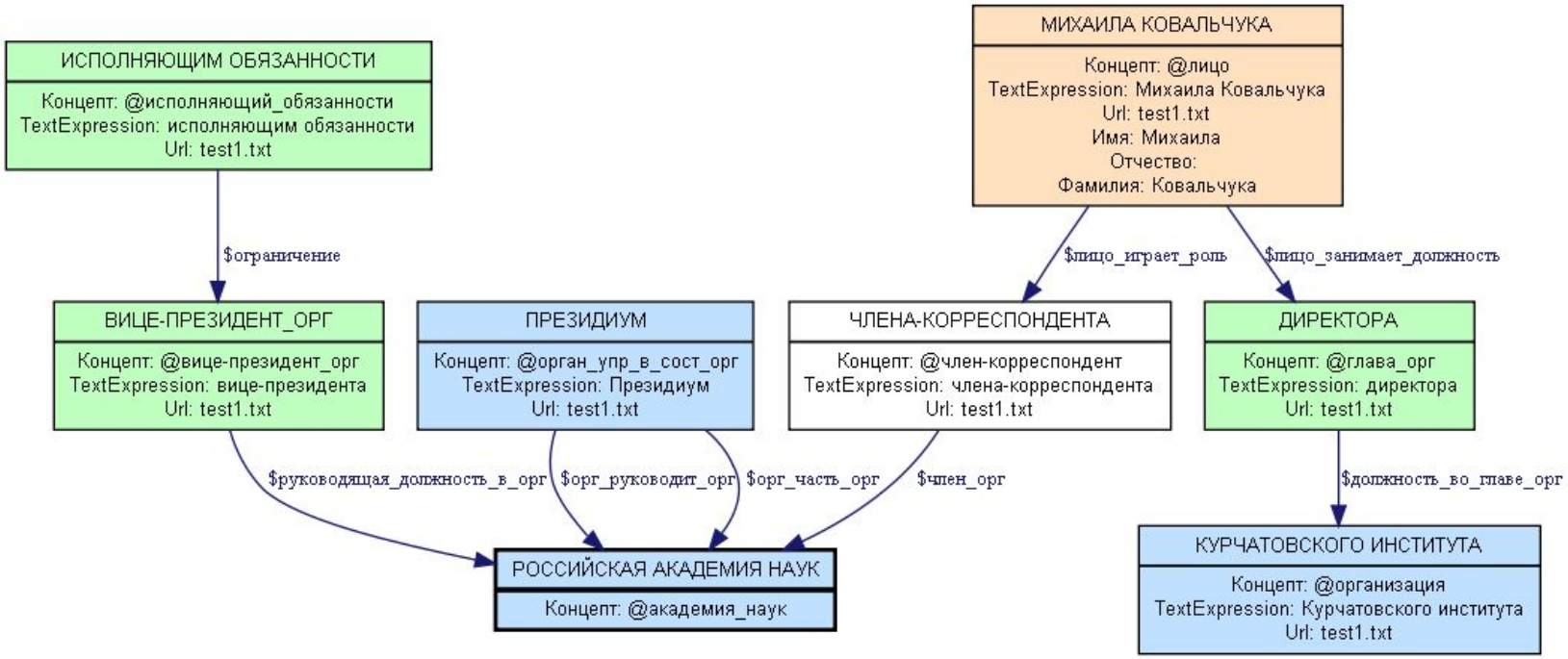
Результаты извлечения информации

Полученные результаты могут использоваться

- непосредственно — система выводит новые факты, распределенные по набору текстов, обеспечивает способ их визуализации
- в качестве исходных данных для систем Data Mining — данные теперь структурированы
- в качестве исходных данных для подсистемы индексирования — это даст новые возможности локального поиска

Пример факта

Президиум Российской академии наук решил назначить члена-корреспондента РАН, директора Курчатовского института Михаила Ковальчука исполняющим обязанности вице-президента РАН.



Установление кореферентности номинаций экземпляров

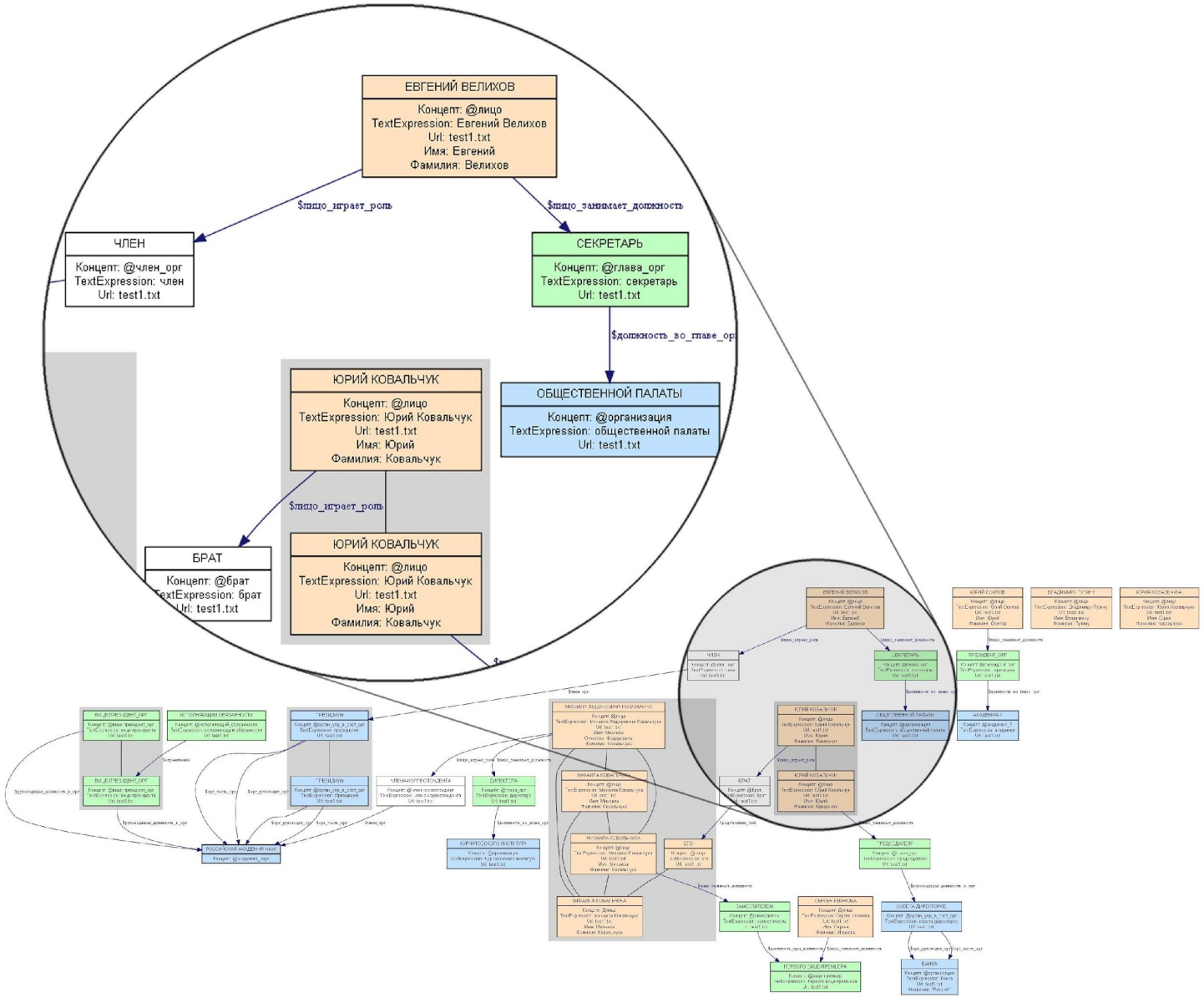
- Разрешение местоименной анафоры
- Установление тождества между номинациями экземпляров из одного текста

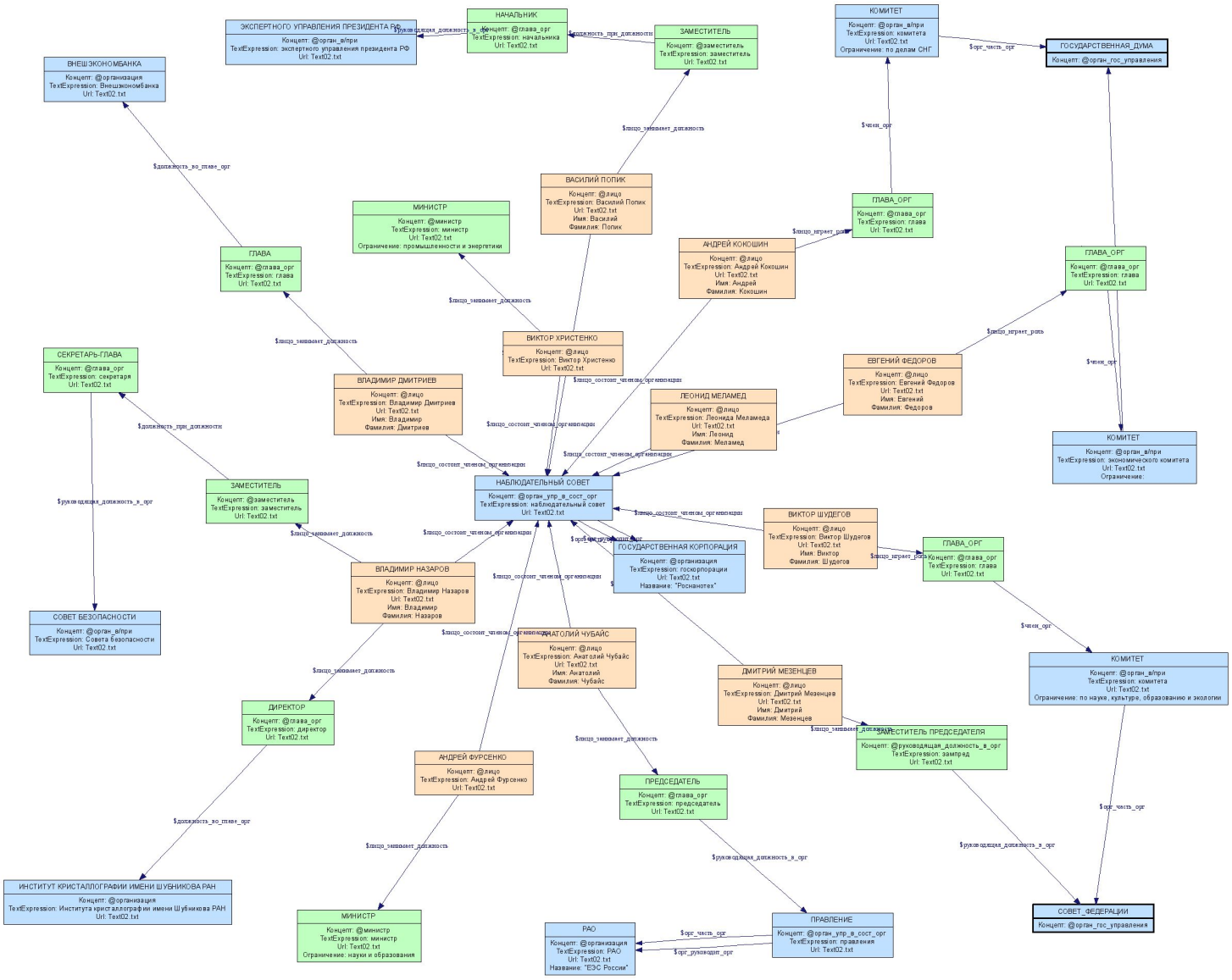
Иллюстрация ⇒

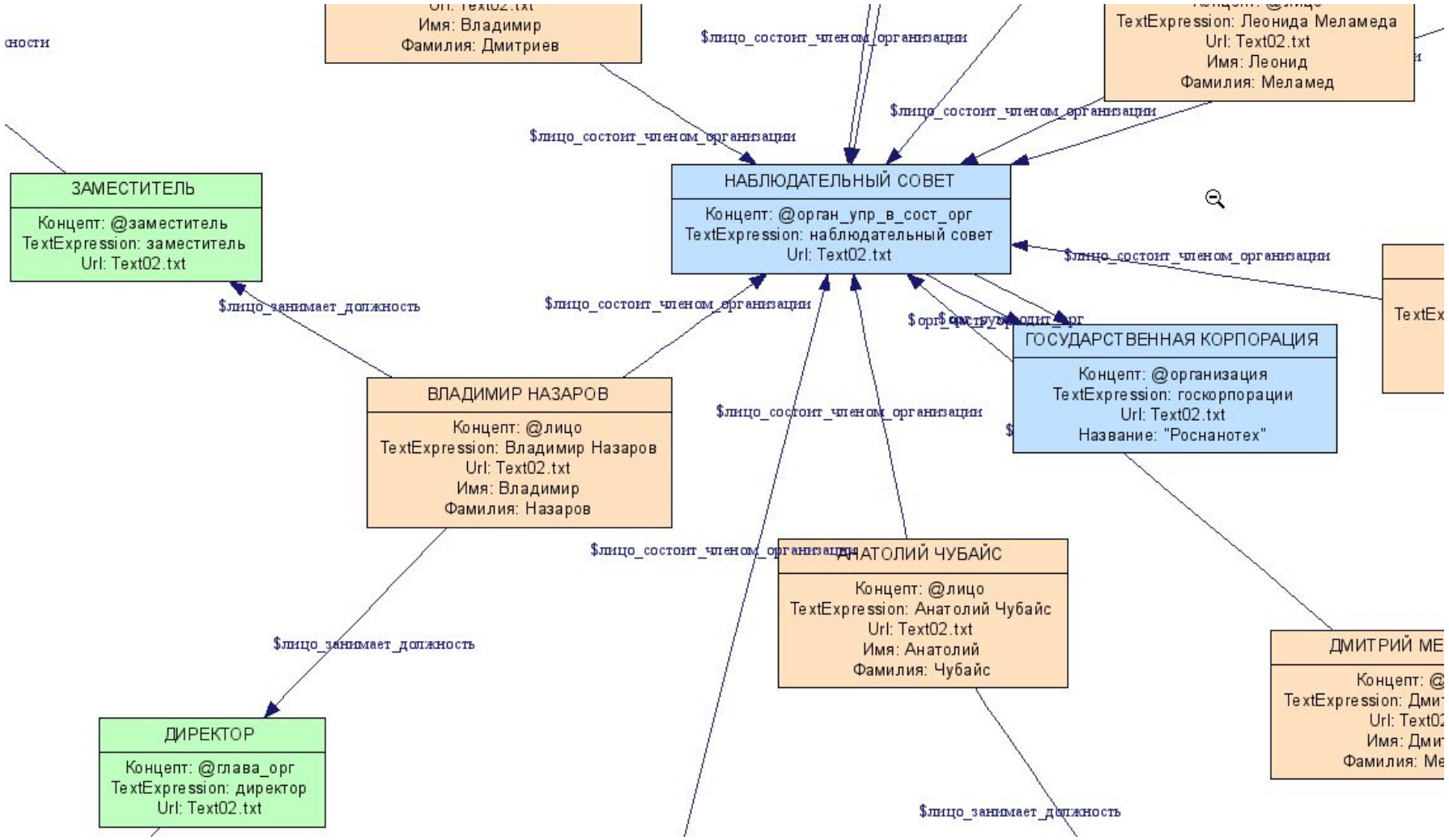
То, что 28-летний Борис Ковальчук будет назначен на эту должность, вопрос практически решенный, и его кандидатура проходит процедуру формального согласования в спецслужбах.

Его отец, Юрий Ковальчук, почетный консул Таиланда в Санкт-Петербурге, в 1996 году наряду с Владимиром Путиным и нынешним министром образования Андреем Фурсенко выступил соучредителем дачного кооператива "Озеро", а в 2000 году создал и возглавил центр стратегических разработок "Северо-запад".

Примеры

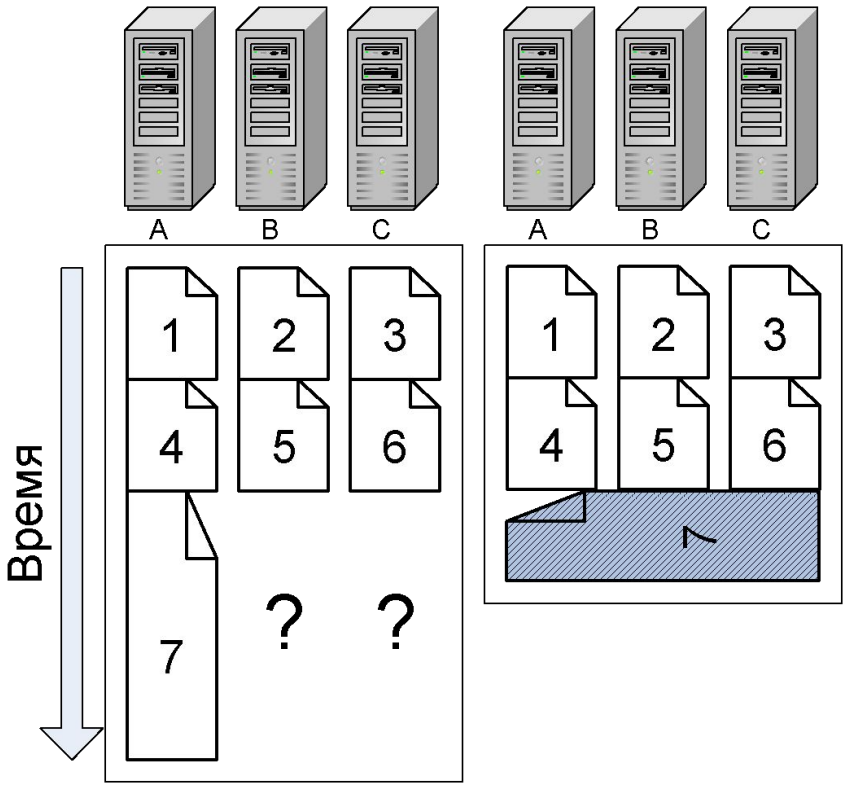

ЕВГЕНИЙ ВЕЛИХОВ
Концепт: @лицо
TextExpression: Евгений Велихов
Url: test1.txt
Имя: Евгений
Фамилия: Велихов
$лицо_играет_роль
$лицо_занимает_должность
ЧЛЕН
Концепт: @член_орг
TextExpression: член
Url: test1.txt
СЕКРЕТАРЬ
Концепт: @глава_орг
TextExpression: секретарь
Url: test1.txt
$должность_во_главе_орг
ЮРИЙ КОВАЛЬЧУК
Концепт: @лицо
TextExpression: Юрий Ковальчук
Url: test1.txt
Имя: Юрий
Фамилия: Ковальчук
ОБЩЕСТВЕННОЙ ПАЛАТЫ
Концепт: @организация
TextExpression: общественной палаты
Url: test1.txt
$лицо_играет_роль
БРАТ
Концепт: @брат
TextExpression: брат
Url: test1.txt
ЮРИЙ КОВАЛЬЧУК
Концепт: @лицо
TextExpression: Юрий Ковальчук
Url: test1.txt
Имя: Юрий
Фамилия: Ковальчук


ЭКСПЕРТНОГО УПРАВЛЕНИЯ ПРЕЗИДЕНТА РФ
Концепт: @орган_в/при
TextExpression: экспертного управления президента РФ
Url: Text02.txt
НАЧАЛЬНИК
Концепт: @глава_орг
TextExpression: начальника
Url: Text02.txt
ЗАМЕСТИТЕЛЬ
Концепт: @заместитель
TextExpression: заместитель
Url: Text02.txt
КОМИТЕТ
Концепт: @орган_в/при
TextExpression: комитета
Url: Text02.txt
Ограничение: по делам СНГ
ГОСУДАРСТВЕННАЯ_ДУМА
Концепт: @орган_гос_управления
ВНЕШЭКОНОМБАНКА
Концепт: @организация
TextExpression: Внешэкономбанка
Url: Text02.txt
ВАСИЛИЙ ПОПИК
Концепт: @лицо
TextExpression: Василий Попик
Url: Text02.txt
Имя: Василий
Фамилия: Попик
МИНИСТР
Концепт: @министр
TextExpression: министр
Url: Text02.txt
Ограничение: промышленности и энергетики
ГЛАВА_ОРГ
Концепт: @глава_орг
TextExpression: глава
Url: Text02.txt
АНДРЕЙ КОКОШИН
Концепт: @лицо
TextExpression: Андрей Кокошин
Url: Text02.txt
Имя: Андрей
Фамилия: Кокошин
ГЛАВА
Концепт: @глава_орг
TextExpression: глава
Url: Text02.txt
ВИКТОР ХРИСТЕНКО
Концепт: @лицо
TextExpression: Виктор Христенко
Url: Text02.txt
ЕВГЕНИЙ ФЕДОРОВ
Концепт: @лицо
TextExpression: Евгений Федоров
Url: Text02.txt
Имя: Евгений
Фамилия: Федоров
СЕКРЕТАРЬ-ГЛАВА
Концепт: @глава_орг
TextExpression: секретаря
Url: Text02.txt
ВЛАДИМИР ДМИТРИЕВ
Концепт: @лицо
TextExpression: Владимир Дмитриев
Url: Text02.txt
Имя: Владимир
Фамилия: Дмитриев
ЛЕОНИД МЕЛАМЕД
Концепт: @лицо
TextExpression: Леонида Меламеда
Url: Text02.txt
Имя: Леонид
Фамилия: Меламед
КОМИТЕТ
Концепт: @орган_в/при
TextExpression: экономического комитета
Url: Text02.txt
Ограничение:
ЗАМЕСТИТЕЛЬ
Концепт: @заместитель
TextExpression: заместитель
Url: Text02.txt
НАБЛЮДАТЕЛЬНЫЙ СОВЕТ
Концепт: @орган_упр_в_сост_орг
TextExpression: наблюдательный совет
Url: Text02.txt
ВИКТОР ШУДЕГОВ
Концепт: @лицо
TextExpression: Виктор Шудегов
Url: Text02.txt
Имя: Виктор
Фамилия: Шудегов
ГЛАВА_ОРГ
Концепт: @глава_орг
TextExpression: глава
Url: Text02.txt
ГОСУДАРСТВЕННАЯ КОРПОРАЦИЯ
Концепт: @организация
TextExpression: госкорпорации
Url: Text02.txt
Название: "Роснанотех"
ВЛАДИМИР НАЗАРОВ
Концепт: @лицо
TextExpression: Владимир Назаров
Url: Text02.txt
Имя: Владимир
Фамилия: Назаров
СОВЕТ БЕЗОПАСНОСТИ
Концепт: @орган_в/при
TextExpression: Совета безопасности
Url: Text02.txt
АНАТОЛИЙ ЧУБАЙС
Концепт: @лицо
TextExpression: Анатолий Чубайс
Url: Text02.txt
Имя: Анатолий
Фамилия: Чубайс
ДМИТРИЙ МЕЗЕНЦЕВ
Концепт: @лицо
TextExpression: Дмитрий Мезенцев
Url: Text02.txt
Имя: Дмитрий
Фамилия: Мезенцев
КОМИТЕТ
Концепт: @орган_в/при
TextExpression: комитета
Url: Text02.txt
Ограничение: по науке, культуре, образованию и экологии
ДИРЕКТОР
Концепт: @глава_орг
TextExpression: директор
Url: Text02.txt
АНДРЕЙ ФУРСЕНКО
Концепт: @лицо
TextExpression: Андрей Фурсенко
Url: Text02.txt
ПРЕДСЕДАТЕЛЬ
Концепт: @глава_орг
TextExpression: председатель
Url: Text02.txt
ЗАМЕСТИТЕЛЬ ПРЕДСЕДАТЕЛЯ
Концепт: @руководящая_должность_в_орг
TextExpression: зампред
Url: Text02.txt
ИНСТИТУТ КРИСТАЛЛОГРАФИИ ИМЕНИ ШУБНИКОВА РАН
Концепт: @организация
TextExpression: Института кристаллографии имени Шубникова РАН
Url: Text02.txt
МИНИСТР
Концепт: @министр
TextExpression: министр
Url: Text02.txt
Ограничение: науки и образования
РАО
Концепт: @организация
TextExpression: РАО
Url: Text02.txt
Название: "ЕЭС России"
ПРАВЛЕНИЕ
Концепт: @орган_упр_в_сост_орг
TextExpression: правления
Url: Text02.txt
СОВЕТ_ФЕДЕРАЦИИ
Концепт: @орган_гос_управления
$руководящая_должность_в_орг
$должность_при_должности
$орг_часть_орг
$должность_во_главе_орг
$лицо_занимает_должность
$член_орг
$лицо_играет_роль
$лицо_состоит_членом_организации
$должность_при_должности
$руководящая_должность_в_орг
$лицо_занимает_должность
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
$лицо_играет_роль
$член_орг
$лицо_занимает_должность
$должность_во_главе_орг
$лицо_занимает_должность
$руководящая_должность_в_орг
$орг_часть_орг
$орг_руководит_орг
$руководящая_должность_в_орг
$орг_часть_орг


Имя: Владимир
Фамилия: Дмитриев
$лицо_состоит_членом_организации
TextExpression: Леонида Меламеда
Url: Text02.txt
Имя: Леонид
Фамилия: Меламед
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
ЗАМЕСТИТЕЛЬ
Концепт: @заместитель
TextExpression: заместитель
Url: Text02.txt
НАБЛЮДАТЕЛЬНЫЙ СОВЕТ
Концепт: @орган_упр_в_сост_орг
TextExpression: наблюдательный совет
Url: Text02.txt
$лицо_состоит_членом_организации
$лицо_занимает_должность
$лицо_состоит_членом_организации
ГОСУДАРСТВЕННАЯ КОРПОРАЦИЯ
Концепт: @организация
TextExpression: госкорпорации
Url: Text02.txt
Название: "Роснанотех"
ВЛАДИМИР НАЗАРОВ
Концепт: @лицо
TextExpression: Владимир Назаров
Url: Text02.txt
Имя: Владимир
Фамилия: Назаров
$лицо_состоит_членом_организации
$лицо_состоит_членом_организации
АНАТОЛИЙ ЧУБАЙС
Концепт: @лицо
TextExpression: Анатолий Чубайс
Url: Text02.txt
Имя: Анатолий
Фамилия: Чубайс
$лицо_занимает_должность
ДИРЕКТОР
Концепт: @глава_орг
TextExpression: директор
Url: Text02.txt
$лицо_занимает_должность

ИСИДА-Т: Параллельная обработка данных

- Параллелизм на уровне документов для высокой производительности и снижения накладных расходов
- Разбиение документов для балансировки нагрузки
- Выделение сервисных узлов для выполнения отдельных функций по необходимости



Параллельные вычисления

Параллельность в системе определяется спецификой конкретной задачи поиска и анализа информации. Выделяются следующие типы параллелизма:

- **по данным** (требуется обрабатывать независимые документы: индексация, извлечение информации…)
- **по задачам** (задачи загрузки документов, их индексации, каталогизации и поиска, работы с ресурсами знаний могут осуществляться независимо друг от друга)
- **по пользователям** (требуется обеспечить распределенную обработку запросов различных пользователей)
- **алгоритмический параллелизм** (некоторые алгоритмы, например вычисления прямого и обратного индекса, могут быть разбиты на параллельные блоки и исполняться на разных узлах)