
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> ...*

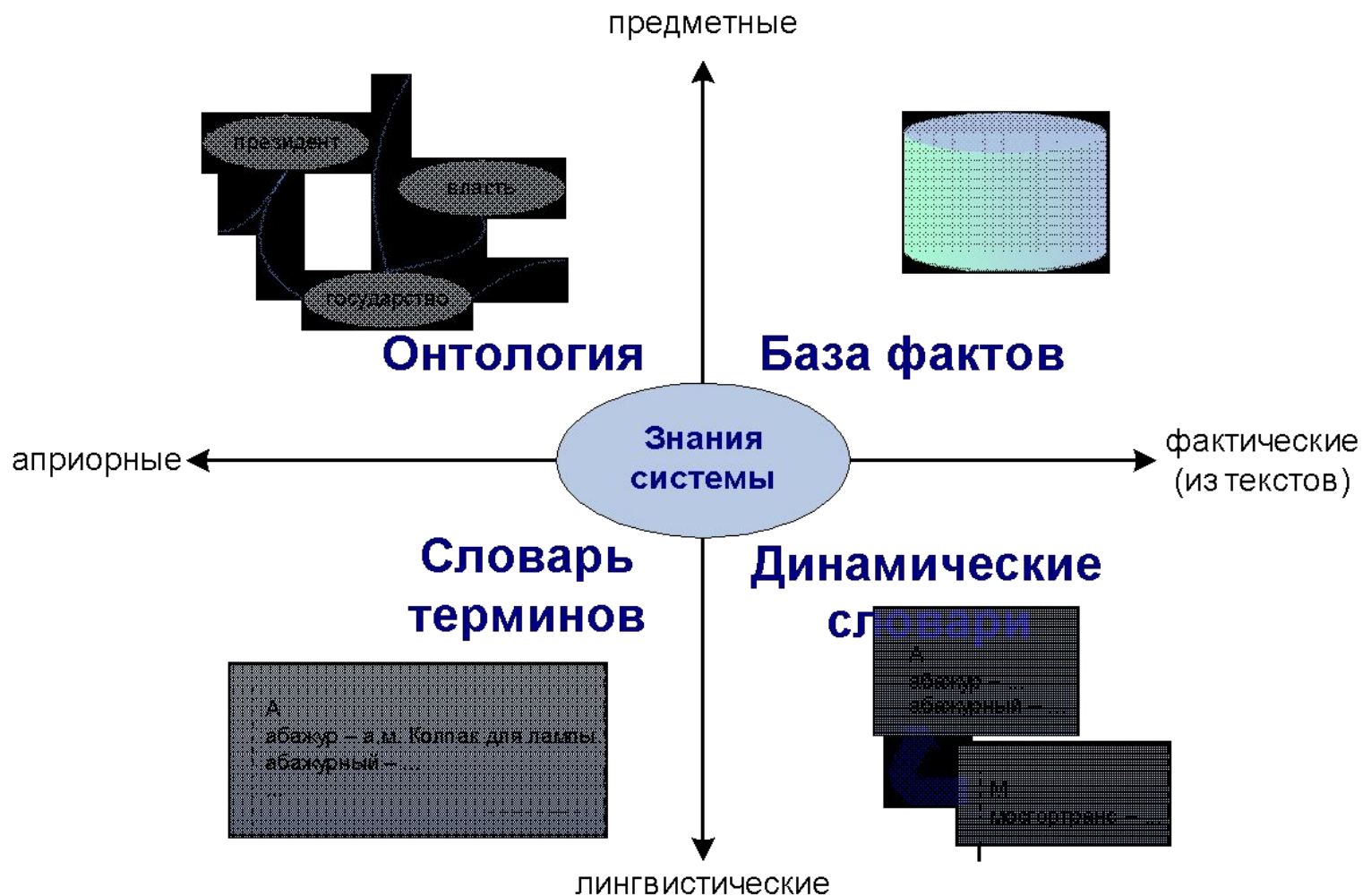
Приложения технологии извлечения информации

- 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
-

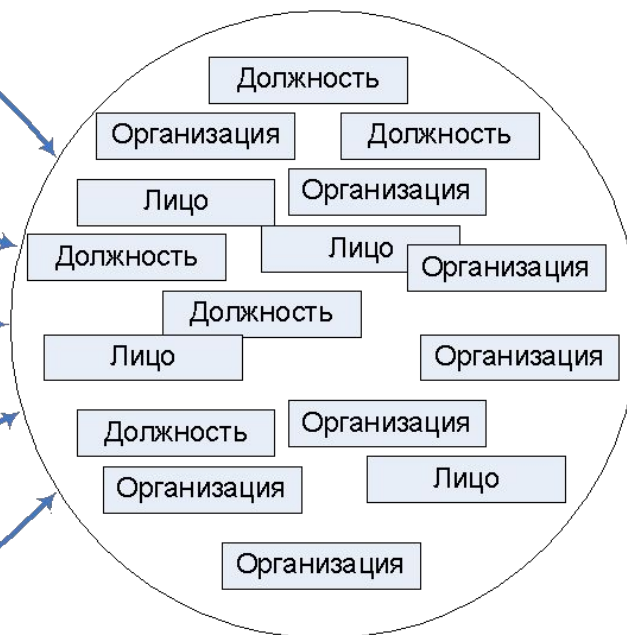
Извлечение информации в «слабом» СМЫСЛЕ

- 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

Правила построения текстовых объектов

- Ковальчук Михаил Федорович* родился 5 января 1948 г. в Черниговской области Украинской ССР.[...] Прошел путь от транспортного рабочего, водителя до генерального директора одного из крупнейших автотранспортных предприятий Санкт-Петербурга - АОЗТ "Трансэк". Возглавляет предприятие с 1978 года. [...] Женат, имеет двоих сыновей.
- Об этом "ДП" сообщил генеральный директор ЗАО "Трансэк" Михаил Ковальчук. //31 июля 2003
- [...] - говорит исполнительный директор транспортной компании ЗАО «Трансэк» Игорь Ковальчук. //06.10.04
- Вчера член-корреспондент РАН Михаил Ковальчук был назначен на должность директора Российского научного центра "Курчатовский институт". //04.02.05
- [...] по словам генерального директора транспортной компании "Трансэк" Игоря Ковальчука, пробок не наблюдается. // 24 апреля 2006 г.



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	

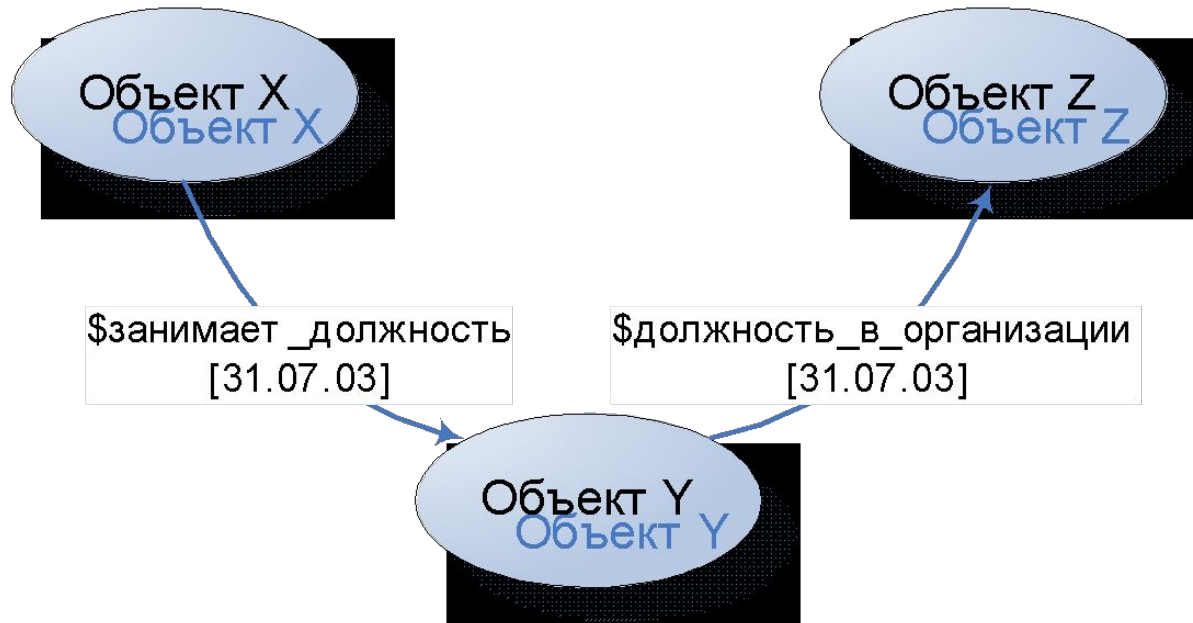
Тип:	@должность		
Подтип:	@генеральный_директор		

			...
--	--	--	-----

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**.

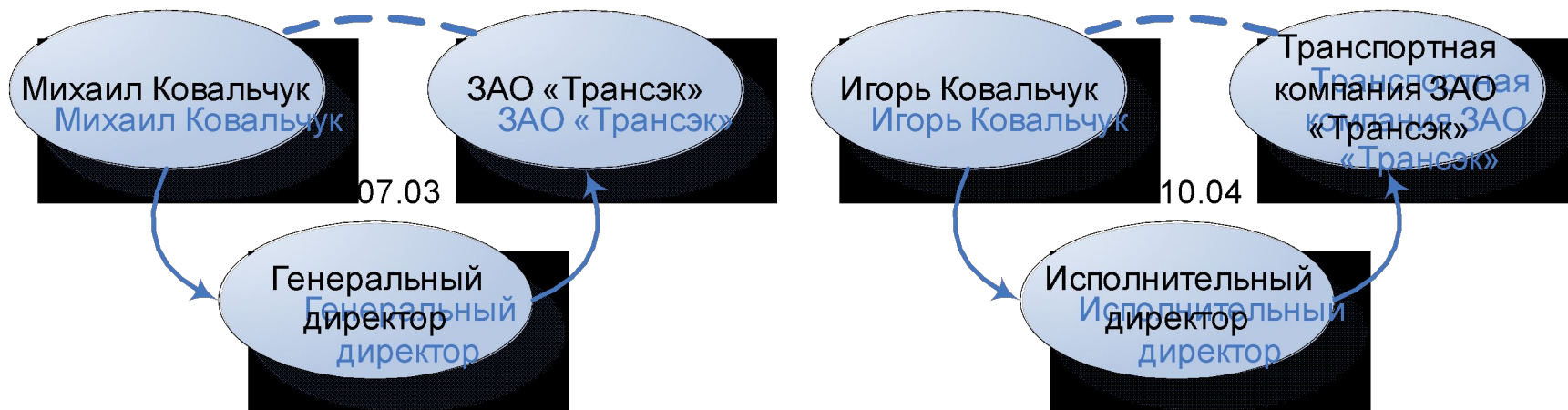
Построение текстовых фактов

- 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]

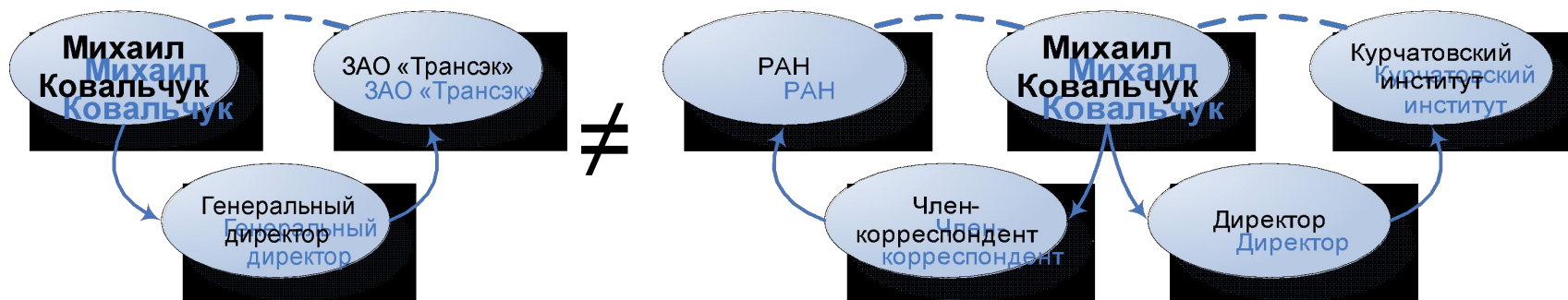
Установление корелферентности (примеры)

- 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

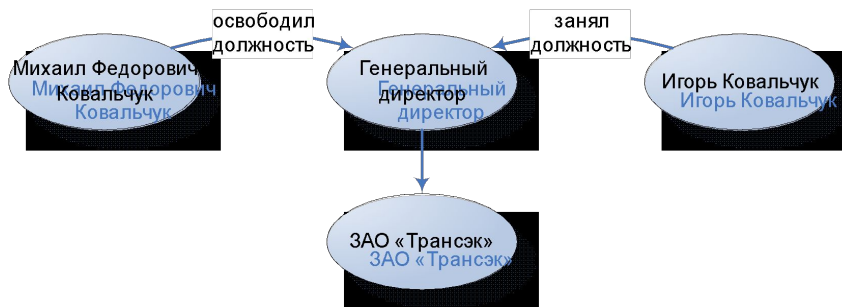
- Устанавливается, что существуют два разных лица с именем *Михаил Ковальчук*:



Вывод новых фактов

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

- «Смена лиц, занимающих должность»

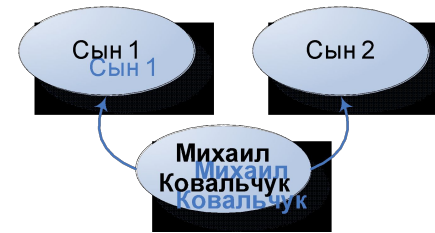


- «Смена должностей лица»



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:
- ...



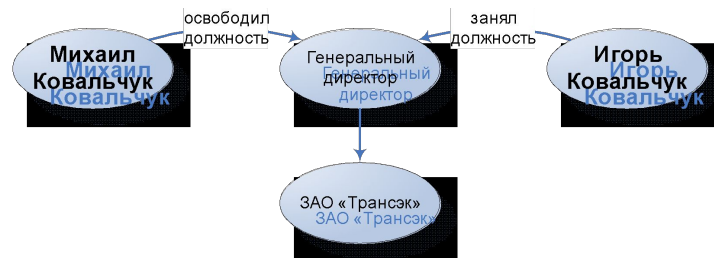
...	
Пол:	мужской
Отчество:	Михайлович
Фамилия:	Ковальчук
...	



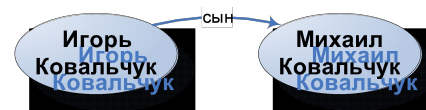
Построение гипотез...

(окончание)

- Система ранее вывела факт:
- Предположим, в знаниях системы о мире есть фрагмент, который позволяет строить гипотезы — например, такого рода:
- Строится гипотеза:
- Для подтверждения или опровержения этой гипотезы у системы пока нет данных. Но они могут появиться по мере поступления новых текстов.



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

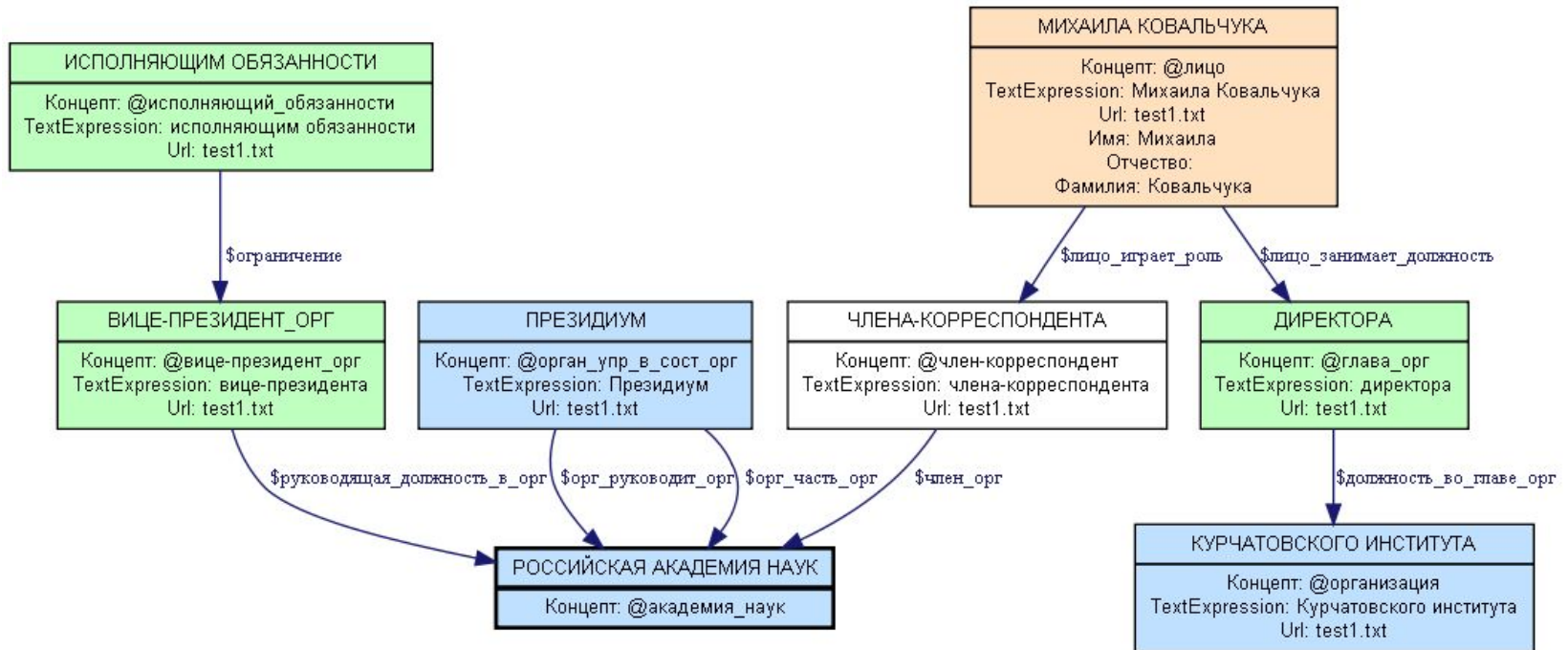


Результаты извлечения информации

- 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

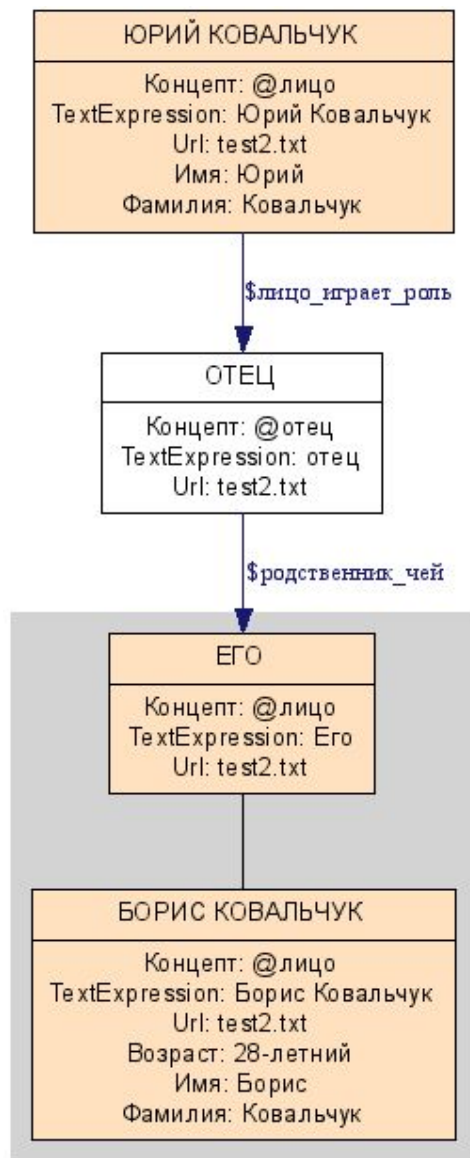
Президиум Российской академии наук решил назначить члена-корреспондента РАН, директора Курчатовского института Михаила Ковальчука исполняющим обязанности вице-президента РАН.



Установление кореферентности номинаций экземпляров

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

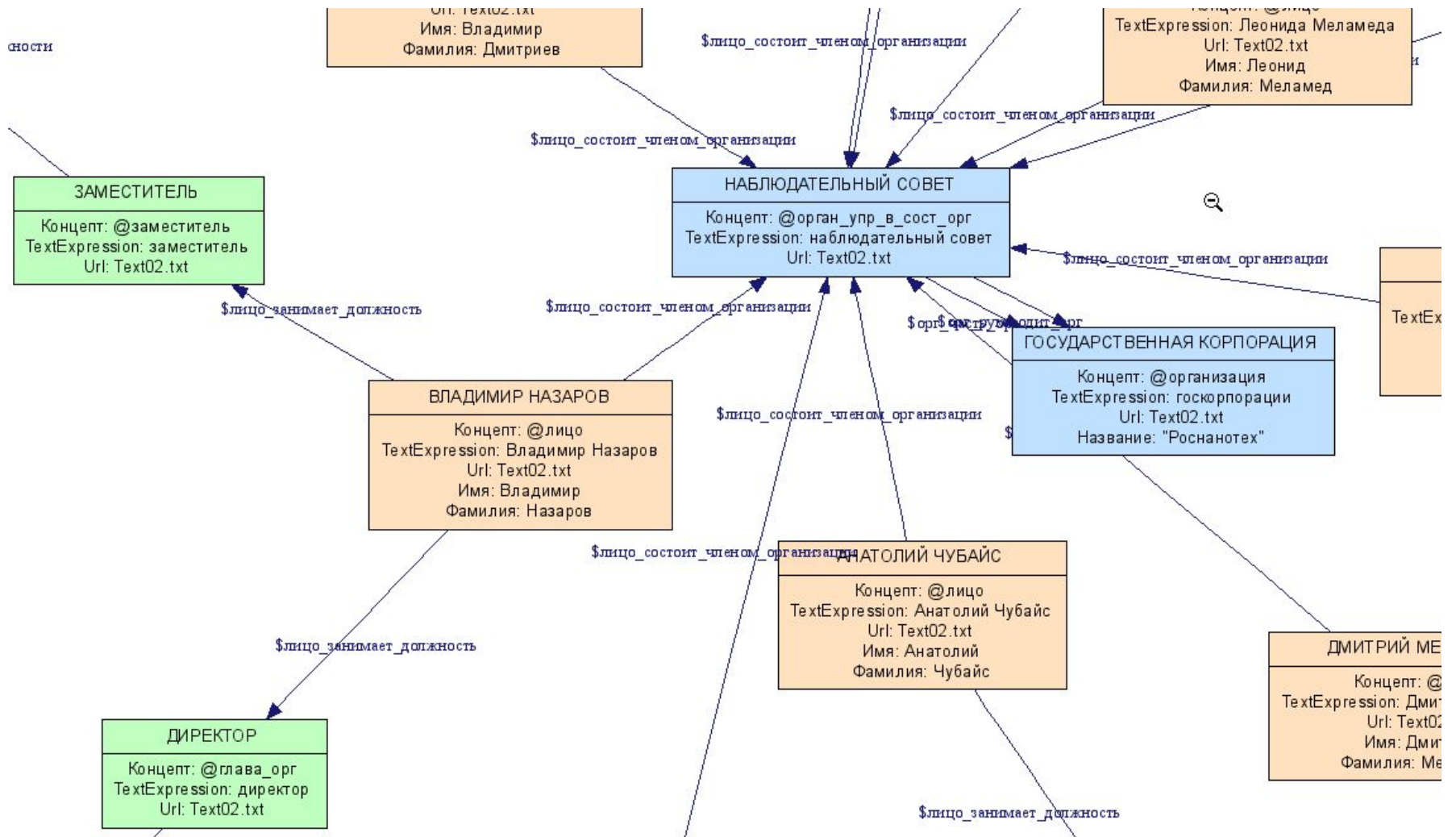
Иллюстрация ⇒



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".

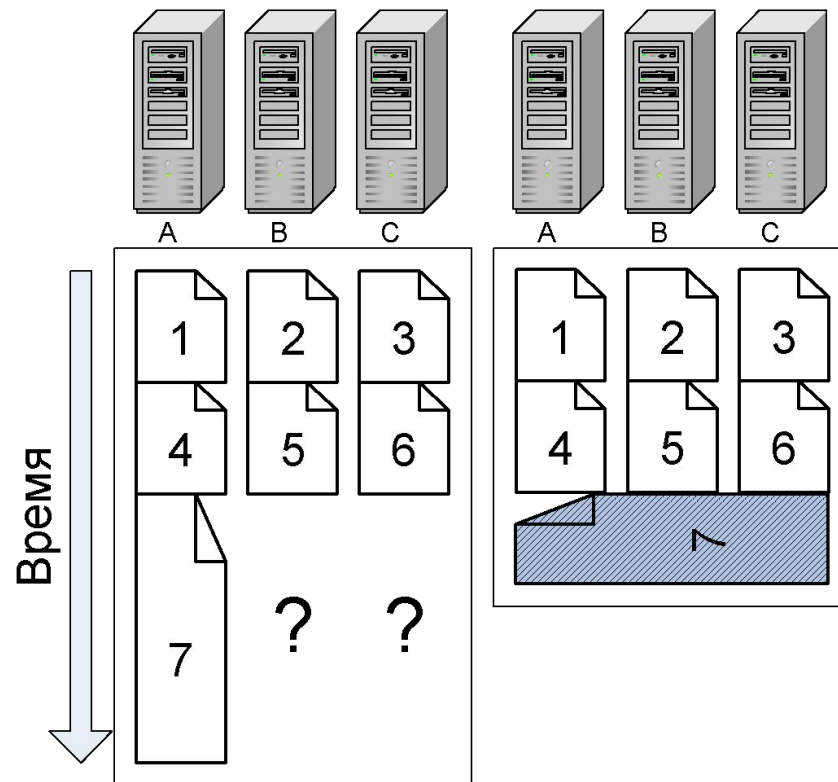
Примеры



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)