



МГУ им. М.В.Ломоносова  
Научно-исследовательский  
вычислительный центр



АНО Центр  
информационных  
исследований

---

Лукашевич Н.В., Добров Б.В.

# Операции с онтологиями: сопоставление, порождение, наращивание, подрезание

# Коллектив



1994 – н/в АНО Центр информационных исследований  
(АНО ЦИИ)

1994 – 1997 Институт США и Канады РАН

1997 – н/в Научно-исследовательский  
вычислительный центр  
МГУ им.М.В.Ломоносова

Университетская информационная система  
РОССИЯ (**УИС РОССИЯ**, [uisrussia.msu.ru](http://uisrussia.msu.ru)):

три миллиона документов (нормативные акты,  
пресса, экономическая статистика)

# Лингвистические ресурсы для автоматической обработки текстовых коллекций: особенности



- ❖ **Наш опыт: развитие ресурсов для задач информационного поиска с 1994 года**
- ❖ **Большой объем: тысячи слов и словосочетаний**
- ❖ **Модель описания знаний о языке и мире должна быть:**
  - ❖ «легкая»,
  - ❖ полезная в широком круге приложений
  - ❖ тестирование ресурса в приложениях
- ❖ **Лингвистические онтологии (тезаурусы)**
  - ❖ **Тезаурус Русского языка РуТез (52 тыс. понятий)**
  - ❖ **Онтология по естественным наукам и технологиям (ОЕНТ) (55 тыс. понятий)**
  - ❖ **Тезаурус (лингв. онтология) по банковской деятельности**
  - ❖ **и др.**

# Клиенты, проекты



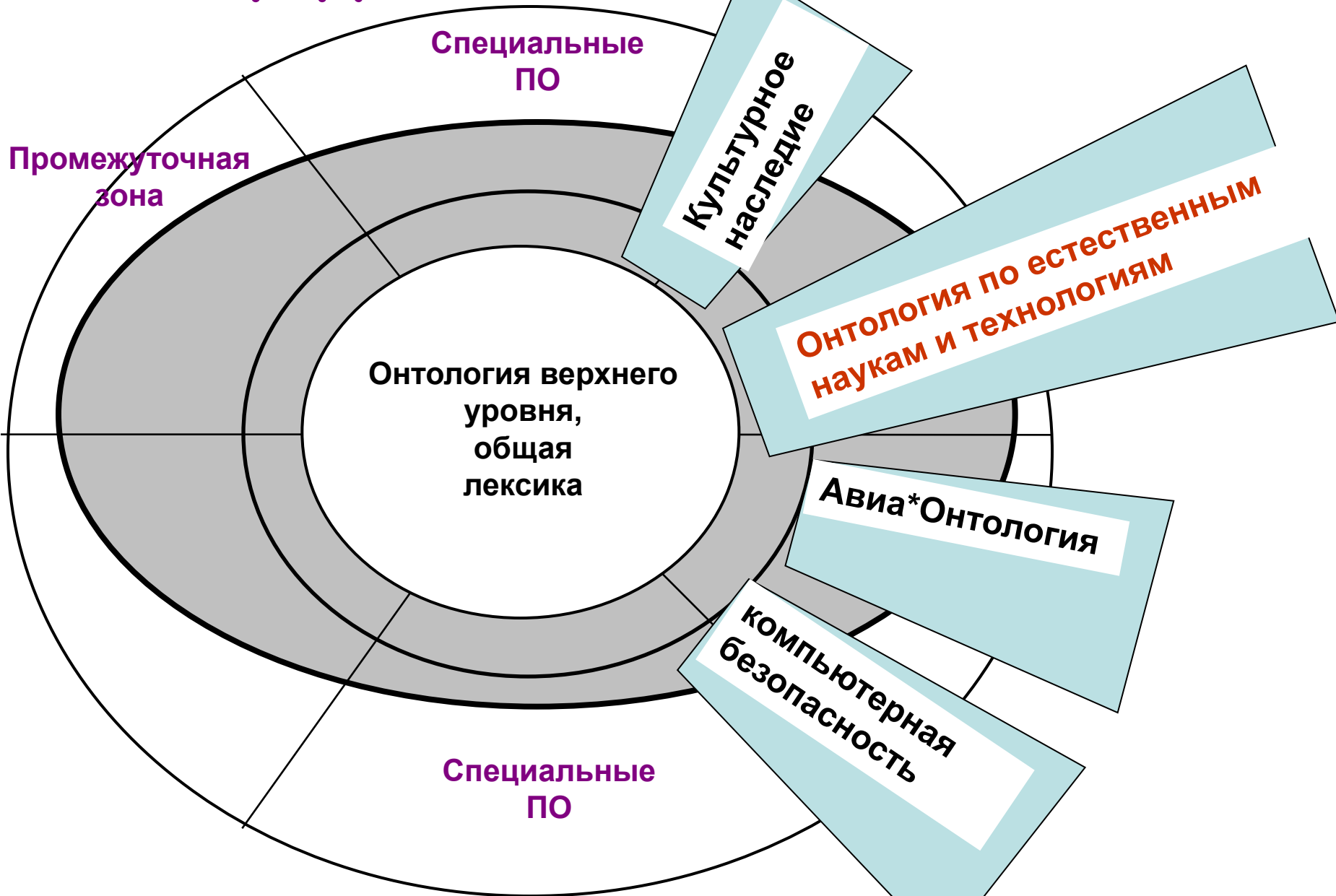
- **Банк России (2006 – н/в)**
- **Рамблер (2007– н/в)**
- **НПП Гарант-Сервис (2002 – н/в)**
- **НИИ Восход для ЦИК РФ (1997 -- н/в)**
- **в/ч 43753 (2000 – н/в)**
  
- **Аппарат Государственной Думы ФС РФ (1999 -- н/в)**
- **ИК «Кодекс» для УОПИ ФСО РФ (2007 – 2008)**
- **Счетная палата (2003)**
- **Министерство образования; ГУМЦ «Базис» (2003, 04)**
- **ИППИ РАН для Управления спецпрограмм (1996)**
- **«Гранит-Центр» (2006), НИЦ «Квант» (2003),  
НТЦ «Атлас» (2001)**



# План презентации

- **Некоторые вопросы использования существующих онтологий**
  - Простые vs. сложные предметные области,
  - Определение границы предметной области,
  - Соотношение «соседних» предметных областей
  - Выгрузка из существующей онтологии нужных фрагментов
- **Оценка качества сопоставления онтологий**
  - семинар по оценке методов сопоставления онтологий ОАЕИ-2009

# Развитие Тезауруса РуТез в сферу специальных областей



# Сложные vs. простые предметные области



- **Простые предметные области**
  - Четкие границы,
  - Границы определяются физическими границами, конкретным процессом (производство, услуги)
  - Ясное назначение сущностей
  
- **Сложные предметные области**
  - Расплывчатые границы,
  - Значимость текстовых документов,
  - Сущности в разных ролях и функциях

# Сложные области: определение границ



- **Междисциплинарность**
  - Государственный финансовый контроль (экономика + право + финансы)
  - Борьба с терроризмом (уголовное право + международное право + государственное право ...)
- **Два подразделения предметной области**
  - Центр предметной области
  - Необходимые разделы из других предметных областей



# Границы области: Государственный финансовый контроль

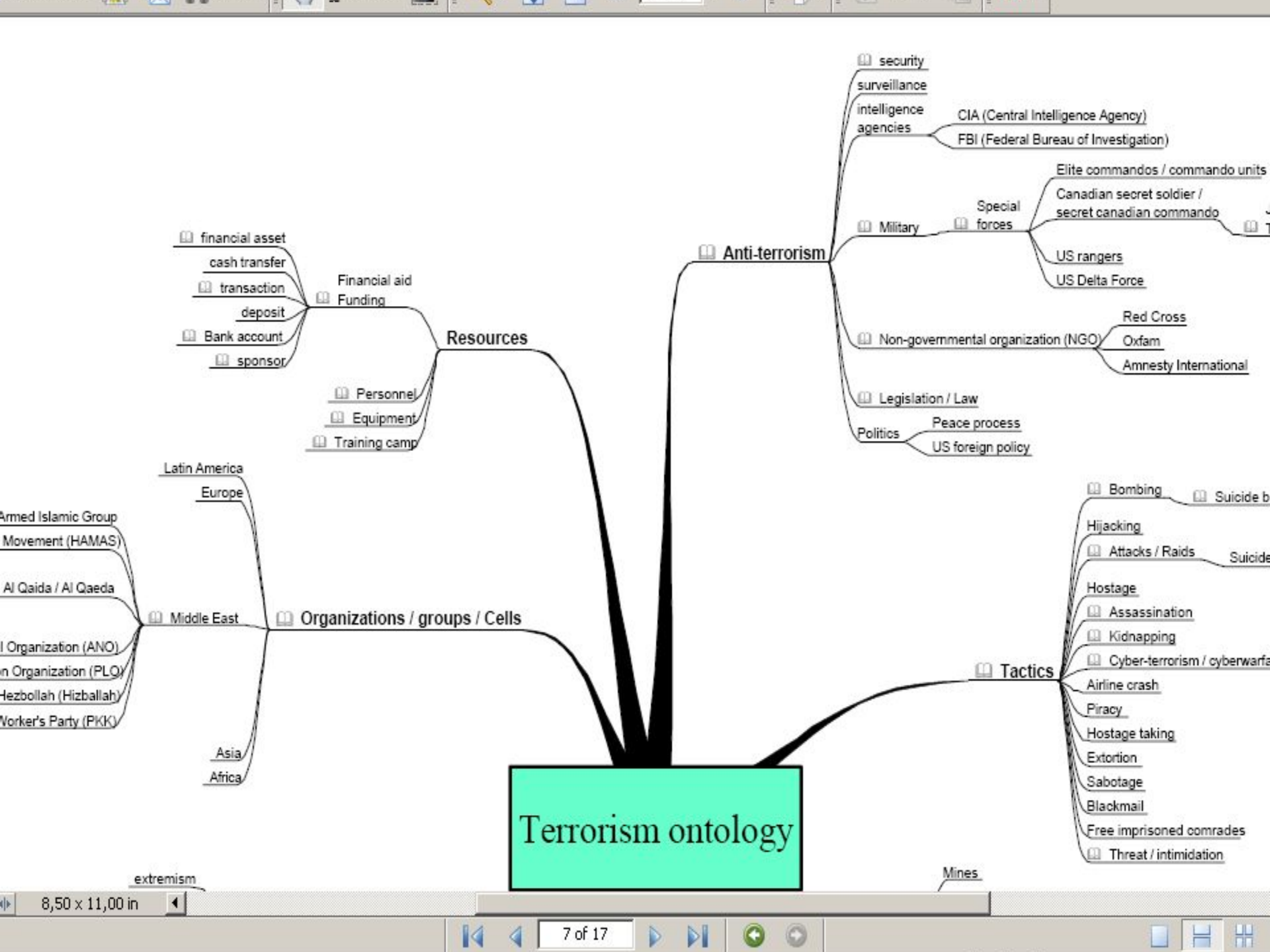


- Термины относящиеся к
  - этапам, процедурам, участникам процесса государственного финансового контроля;
  - к бюджетной системе и бюджетному процессу;
  - к области приобретения, использования и распоряжения государственной собственностью;
  - проверяемым типам деятельности, и основные типы проверяемых документов;
  - термины, описывающие основные организационно-правовые формы организаций в Российской Федерации.

# Границы области: борьба с терроризмом



- **Центр предметной области**
  - Террористические акты
  - Профилактика, борьба с терроризмом и т.п.
- **Вспомогательные разделы**
  - Населенные пункты,
  - Оружие и взрывчатые вещества,
  - Транспорт,
  - Финансовые расчеты,
  - Идеология и религия и др.
- **Казалось бы: торжество концепции вторичного использования онтологий**



# Terrorism ontology

extremism

Mines

# Проблема: искажение реальности



- Общие понятия, необходимые для предметной области, трактуются как относящиеся к этой предметной области
- Название концепта сохраняется общим, а значение подразумевается относящимся к этой предметной области
  - ЗАКОНОДАТЕЛЬСТВО  
(=антитеррористическое законодательство=),
  - РАЗВЕДКА  
(разведка против террористической деятельности)
- Проблемы при склейке, вторичном использовании онтологий
- Тезаурус по радиационному терроризму  
(Radiological terrorism)

# Пример: искажение реальности

## combatting radiological terrorism

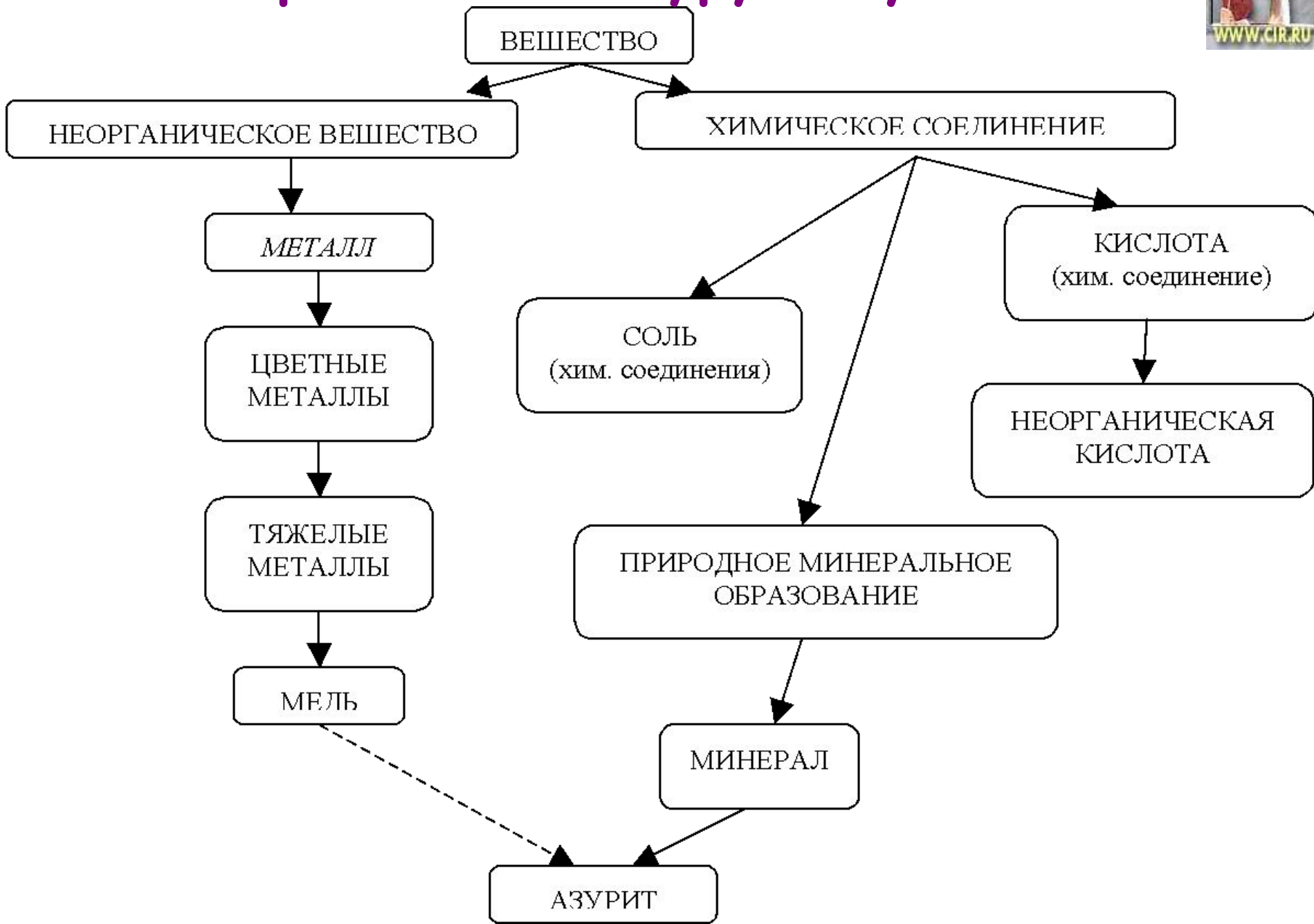
- .antiterrorism
- .counterterrorism
- .intelligence
- ..radiological terrorism requirements
- ...intensive industrial infrastructure
- ...expertise with radiological material
- ...availability of radioactive material
- ..inspection
- ..detection
- .consequence management
- ..medical response
- ...patient decontamination
- ...triage
- ....biodosimetry assessment tool
- ....Andrews lymphocyte nomogram
- ...treatments
- ....Potassium Iodide
- ....Diethylenetriaminepentaacetate
- .....Zn-DTPA
- .....Ca-DTPA
- ....Prussian blue
- ....Filgrastim
- ...neutropenia
- ...pain management
- ...necrosis
- ...reconstructive surgery
- ..public response
- ...evacuation
- ...seeking shelter
- ...removing clothing
- ...washing

# Изменения в описаниях понятий, полученных из тезауруса RuTез



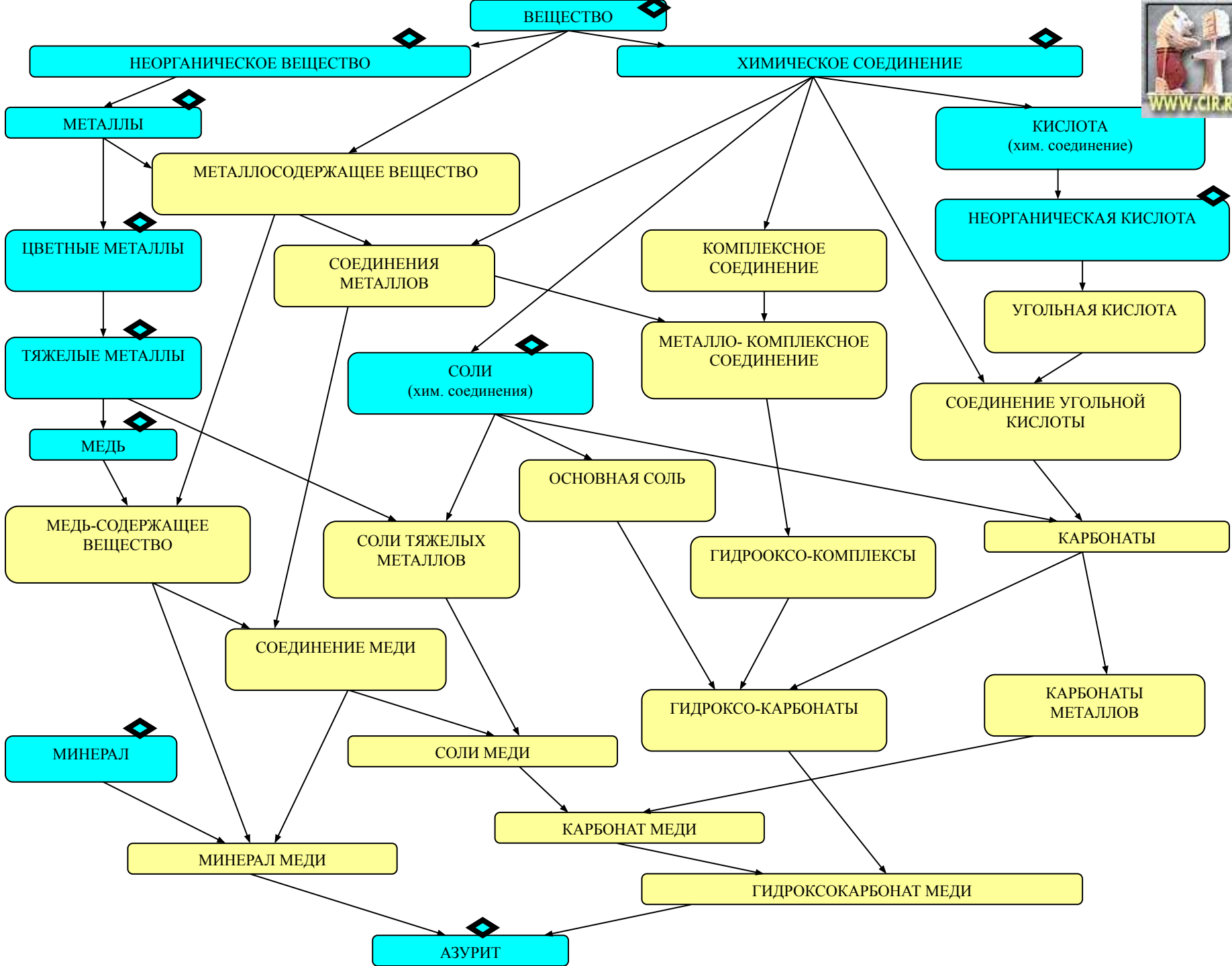
1. Изменение названия понятия;
2. Изменение набора текстовых входов понятия:
3. Изменение отношений между понятиями онтологии-прототипа:
  1. Исчезновение отношений между понятиями онтологии-прототипа;
  2. Появление новых отношений между понятиями онтологии-прототипа;
4. Введение отношений понятий онтологии-прототипа с новыми понятиями:
  1. а. Введение отношений вверх по иерархии;
  2. б. Введение отношений вниз по иерархии

# Фрагмент Тезауруса РуТез





WWW.CIR.RU





# Как эффективно извлечь необходимые концепты и отношения для вторичного использования



- Методы, основанные на специализированном корпусе (Tf\*idf)
- Автоматизированная технология
  - Анализ целевой предметной области,
  - Выявление границ целевой ПО, типов необходимых сущностей
  - Создание рубрикатора, описание рубрик как логических выражений на концептах
  - Выгрузка концептов, попавших в описание рубрик
  - Специализированные корпуса как вспомогательный источник

# Описание рубрик в виде явной логической формулы



## Рубрика

### «Банковские операции и сделки»

[ БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ(E);  
ИНВЕСТИЦИОННЫЙ ФОНД(-,E);  
ВЗАИМОЗАЧЕТ(-,E);  
ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(-,E)

.OR.

[ КРЕДИТНАЯ ОРГАНИЗАЦИЯ(L)

.and.

[ ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(E);  
РИСК (ВОЗМОЖНОСТЬ ОПАСНОСТИ,  
НЕУДАЧИ)(L)

The screenshot shows the 'Рубрика [Банковские операции и сделки]' interface. It displays logical formulas for 'Дизъюнкты' and 'Конъюнкты' with associated weights. Below these are tables for 'Order' and 'Приписка концептов'.

**Дизъюнкты**

Дизъюнкты	Вес	Действие
@[БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ(E);ИНВЕСТИЦИОННЫЙ ФОНД(-,E);ВЗАИМОЗАЧЕТ(-,E);ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(-,E)]	1	Добавить
@[КРЕДИТНАЯ ОРГАНИЗАЦИЯ(L)].and.[ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(E);РИСК (ВОЗМОЖНОСТЬ ОПАСНОСТИ, НЕУДАЧИ)(L)]	1	Изменить

**Конъюнкты**

Конъюнкты	Вес	Действие
@БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ(E);ИНВЕСТИЦИОННЫЙ ФОНД(-,E);ВЗАИМОЗАЧЕТ(-,E);ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(-,E)	1	Добавить
@КРЕДИТНАЯ ОРГАНИЗАЦИЯ(L)	1	Изменить

**Table 1: Order**

Order	Опорные концепты	Знак	Расш.	Вкл.	Подт.
100	БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ	+	E	True	A--
100	ВЗАИМОЗАЧЕТ	--	E	True	M+
100	ДОСТАТОЧНОСТЬ КАПИТАЛА	--	E	True	M+
100	ИНВЕСТИЦИОННЫЙ ФОНД	--	E	True	M+
100	ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ	--	E	True	M+
100	УЧЕТНАЯ СТАВКА	--	E	True	M+
100	УЧЕТНАЯ СТАВКА ЦЕНТРАЛЬНОГО БАНКА	--	E	True	M+

**Table 2: Приписка концептов**

Приписка концептов	Подт.
АВИЗО	
АКТИВНАЯ БАНКОВСКАЯ ОПЕРАЦИЯ	
АУКЦИОН БАНКА РОССИИ	
БАЗОВАЯ ПРОЦЕНТНАЯ СТАВКА	
БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ	
БАНКОВСКАЯ ОПЕРАЦИЯ	
БАНКОВСКАЯ ОПЕРАЦИЯ СО СЧЕТОМ	
БАНКОВСКАЯ СТАВКА	
БАНКОВСКАЯ ТАЙНА	
БАНКОВСКАЯ ЦЕННАЯ БУМАГА	

**Дизъюнкты (second screenshot)**

Дизъюнкты	Вес	Действие
@[БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ(E);ИНВЕСТИЦИОННЫЙ ФОНД(-,E);ВЗАИМОЗАЧЕТ(-,E);ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(-,E)]	1	Добавить
@[КРЕДИТНАЯ ОРГАНИЗАЦИЯ(L)].and.[ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(E);РИСК (ВОЗМОЖНОСТЬ ОПАСНОСТИ, НЕУДАЧИ)(L)]	1	Изменить

**Конъюнкты (second screenshot)**

Конъюнкты	Вес	Действие
@БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ(E);ИНВЕСТИЦИОННЫЙ ФОНД(-,E);ВЗАИМОЗАЧЕТ(-,E);ПЕРЕВОД ДЕНЕЖНЫХ СРЕДСТВ(-,E)	1	Добавить
@КРЕДИТНАЯ ОРГАНИЗАЦИЯ(L)	1	Изменить

**Table 3: Order (second screenshot)**

Order	Опорные концепты	Знак	Расш.	Вкл.	Подт.
100	КРЕДИТНАЯ ОРГАНИЗАЦИЯ	+	L	True	A--

**Table 4: Приписка концептов (second screenshot)**

Приписка концептов	Подт.
АВТОБАНК	
АКЦИОНЕРНЫЙ БАНК	
АЛЬФА-БАНК	
АМЕРИКАНСКИЙ БАНК	
БАНК "ВОЗРОЖДЕНИЕ"	
БАНК (ФИНАНСОВОЕ УЧРЕЖДЕНИЕ)	
БАНК АНГЛИИ	
БАНК МЕЖДУНАРОДНЫХ РАСЧЕТОВ	
БАНК МОСКВЫ	
БАНК РОССИИ	
БАНК С УЧАСТИЕМ ИНОСТРАННОГО КАПИТАЛА	
БАНК-АГЕНТ	
БАНК-КОРРЕСПОНДЕНТ	

# Быстрое описание границ предметной области



## Задачи описания границ

□ связность понятийной сети предметной области

□ по иерархии

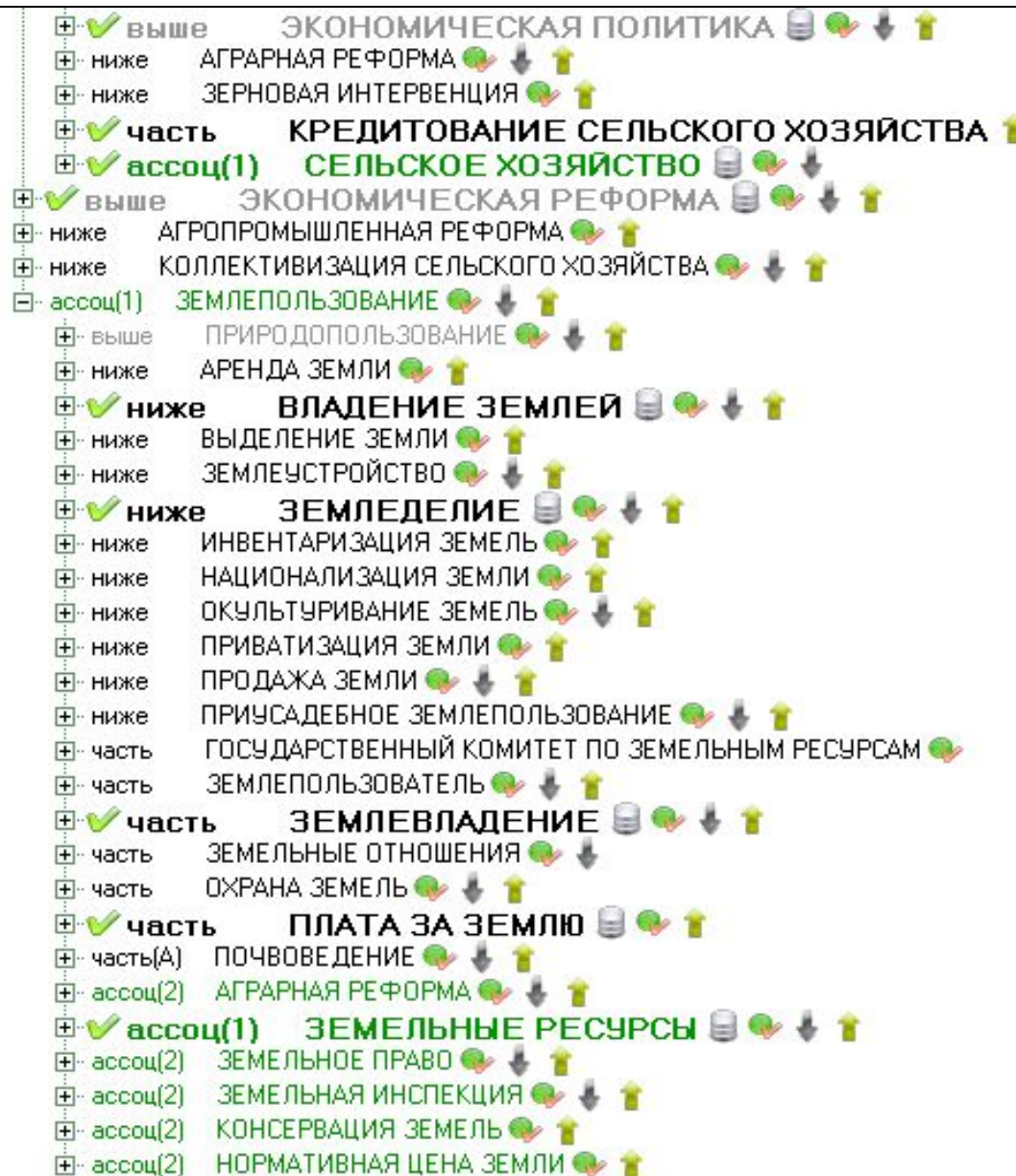
□ по «сестрам»

□ устойчивость границы

□ отсутствие «выбросов»

□ отсутствие «анклавов»

□ замыкание разорванных иерархических связей



# План презентации



- **Некоторые вопросы использования существующих онтологий**
  - Простые vs. сложные предметные области,
  - Определение границы предметной области,
  - Соотношение «соседних» предметных областей
  - Выгрузка из существующей онтологии нужных фрагментов
- **Оценка качества сопоставления онтологий**
  - семинар по оценке методов сопоставления онтологий ОАЕИ-2009

# Семинар ОАЕИ-2009



- **Тестирование методов установления соответствий между единицами онтологий**
- **Тесты проводятся на онтологиях разного уровня формализации (OWL, тезаурусы, рубрикаторы)**
- **Различные типы и меры оценки**
- **4 семинара с 2004 года**
- **5 соревновательных дорожек (11 тестов)**
- **16 участников**

# Трек 1: Базовый (benchmark test)



- **Онтология библиографии**
  - OWL-DL, RDF/XML
  - 33 класса, 60 свойств, 70 экземпляров
- **Тесты серия 1**
  - Сравнение с нерелевантной онтологией (Онтология вина)
  - OWL-DL -> OWL-Lite
- **Тесты серия 2 (замена или отбрасывание)**
  - имена сущностей,
  - комментарии,
  - таксономии,
  - экземпляры
  - свойства

# Трек 1: Базовый (benchmark test) (cont'd)



- **Тесты серия 3**
  - Сопоставление с другими библиографическими онтологиями
- **Лучшие результаты:**
- **Серия 1:**
  - точность – 1, полнота – 1.
- **Серия 2:**
  - точность – 0.97, полнота – 0.86.
- **Серия 3:**
  - точность – 0.84, полнота – 0.81



# Трек 2: Анатомия



- **Сопоставление онтологий**
  - Анатомия человека Института рака
  - Анатомия мышцы
- **61% тривиальных соответствий, т. е. тривиальный уровень результатов**
  - Точность – 0.99, Полнота – 0.60
- **Лучшие результаты:**
  - Точность – 0.95
  - Полнота – 0.77
  - F-мера – 0.855
- **Время работы: 1-20 минут**



# Сопоставление легких (shallow) онтологий: веб-рубрикаторы



- **Системы: Google, Yahoo, Looksmart**
  - Таксономии: отношение subClassOff
  - 300000 категорий в каждом рубрикаторе
  - Моделирование реальной задачи, включающей терминологические проблемы
- **Результаты:**
  - F-меры – 63%
  - Системы обнаружили только 68% положительных соответствий
  - 26% соответствий были найдены всеми участниками
  - 17% отрицательных соответствий были приняты всеми участниками как положительные

# Трек: Библиотека



- **Предметные рубрики библиотек**
  - Библиотека Конгресса США (250 тыс.)
  - Французская национальная библиотека (150 тыс.)
  - Немецкая национальная библиотека (160 тыс.)
- **Информация: синонимы, отношения выше, ниже, ассоциация**
- **Эксперты: 100 тысяч соответствий**
- **Результаты**
  - 1 участник (автоматический перевод)
  - Низкая полнота
  - Проблемы с установлением отношений, отличных от отношений эквивалентности



**Shvaiko P., Euzenat J.:**

# **Ten Challenges for Ontology Matching**

- **Организация масштабного тестирования**
- **Скорость выполнения операций по сопоставлению**
- **Нехватка неявных (background) знаний**
  - **Использование Интернет, предметно-ориентированных текстовых коллекций, онтологий**
- **Выбор и настройка инструмента**
- **Вовлечение пользователя**
- **Объяснение результатов сопоставления**



# Заключение

- Вторичное использование существующих онтологий в практических задачах представляет собой комплексный процесс
- Сложные взаимосвязи между соседними предметными областями
- Проблемы границ, искажение реальности
- Эти проблемы находят свое отражение (и должны учитываться) в процессе практического сопоставления онтологий