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

АНО Центр
информационных
исследований

ГУ НИМЦ
"Базис"

Б.В. Добров, Н.В. Лукашевич,
М.Н. Синицын, В.Н. Шапкин

Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска

(лингвистические и информационные технологии)

Поиск научно-технической информации

- **обеспечение поиска, основанного на знаниях,**
- **использование синонимов,**
- **автоматическое расширение запроса,**
- **автоматический анализ результатов запроса**
- **помощь в интерактивном поиске**

Традиционные средства тематического поиска - информационно-поисковые тезаурусы

- **Основные понятия ПО – дескрипторы**

- **Условные синонимы – аскрипторы**

- **Отношения между дескрипторами:**

  -- ВЫШЕ-НИЖЕ – транзитивно, несимметрично

  -- АССОЦИАЦИЯ – симметрично

  -- Три-четыре уровня иерархии

Традиционные ИП тезаурусы и автоматическая обработка текстов

- **Процесс индексирования базируется на знаниях эксперта**
  - **Удобство для эксперта, относительно небольшая величина**
  - **Дескрипторы нужны для описания основной темы**
  - **Нехватка знаний о понятиях и языке предметной области**
- **Отношения**
  - **Проблема с автоматическим расширением запроса**
  - **Особенно отношение ассоциации**

Семантический поиск в Интернет – Semantic Web: ОНТОЛОГИИ

- **Онтология - это система, состоящая из набора понятий и набора утверждений об этих понятиях, на основе которых можно строить классы, объекты, отношения, функции и теории**

- **Основные компоненты:**
  - **Классы или понятия, примеры**
  - **Отношения, функции**
  - **Аксиомы / правила вывода**

Виды онтологий по составу

1) **Словарь с определениями**
2) **Простая таксономия**
3) **Тезаурус (таксономия с терминами)**
4) **Модель с произвольным набором отношений**
5) **Таксономия и произвольный набор отношений**
6) **Полностью аксиоматизированная теория (фундаментальная онтология)**

Виды онтологий по применению

- **Фундаментальные онтологии**
- **Прикладные онтологии (application ontologies) –**
  легкие онтологии (lightweight ontologies)
  таxономии, ассоциативные тезаурусы
- **Лингвистические онтологии –**
  понятия онтологии связаны со значениями языковых выражений (слов, именных групп и т.п.)

Подходы к описанию отношений при разработке онтологий

- **отношения – произвольный предикат, свойства задаются аксиомами**
- **P(x1,…xn)**
- **Для того, чтобы такая система отношений работала, нужно стабильно находить отношения в разнообразных текстах**
- **Но это проблема!**

Формализация описания области научного знания

- **Цель:**
  обеспечение автоматических процедур тематической обработки и поиска текстов

- **Традиционные информационно-поисковые тезаурусы – недостаточно**

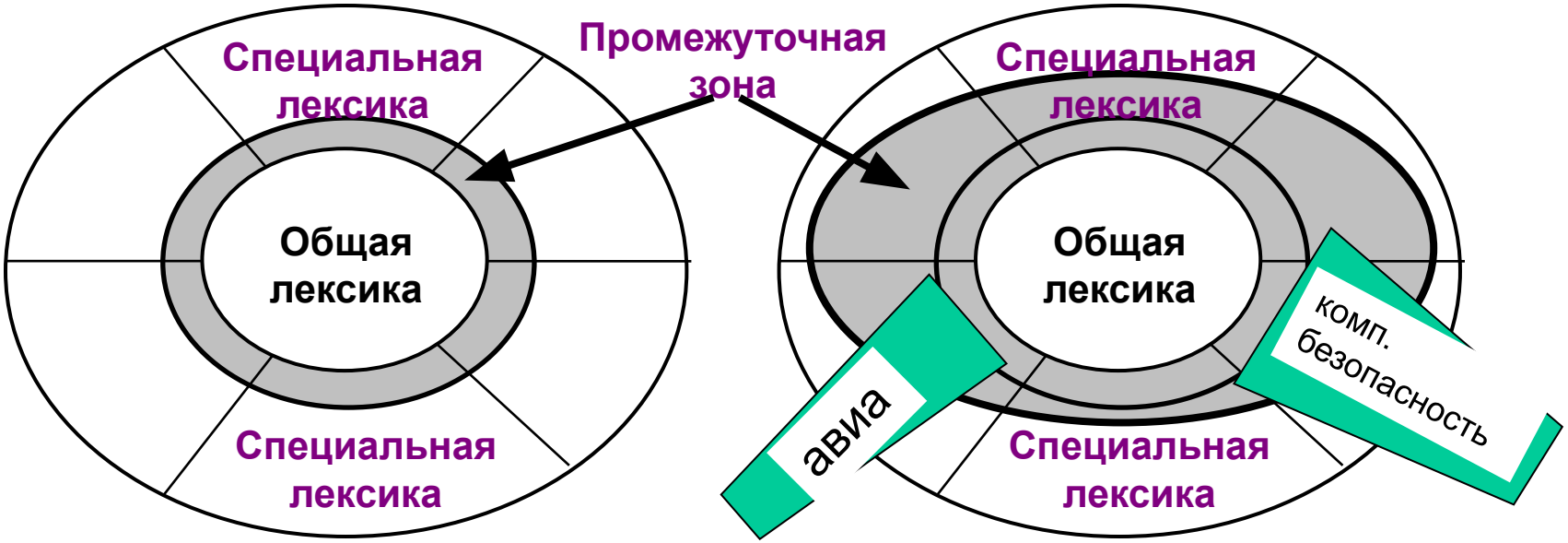
- **Фундаментальные онтологии – невозможно**

Отправная точка

- **Информационно-поисковый тезаурус по общественно-политической тематике РуТез (Общественно-политический тезаурус)**
  - **32 тысячи понятий**
  - **79 тыс. русскоязычных текстовых входов**
  - **80 тыс. англоязычных текстовых входов**

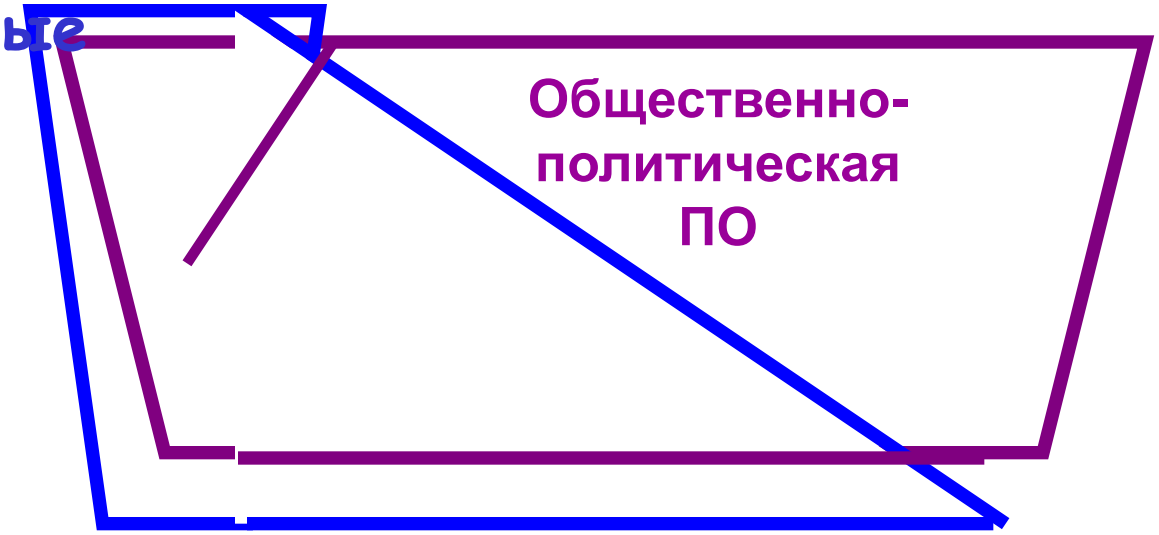
Автоматическая обработка текстов

- **Автоматическое концептуальное индексирование**
- **Автоматическая рубрикация**
- **Автоматическое аннотирование**

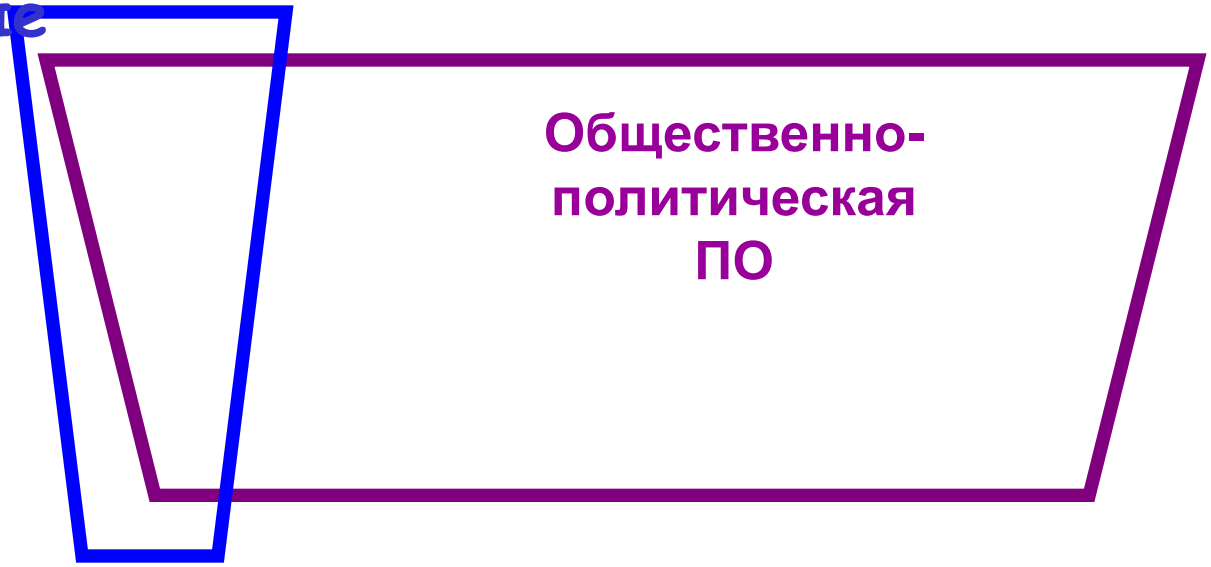


Наука vs Общественно-политическая ПО

Общественные науки



Естественные науки



Лингвистическая онтология по естественным наукам: сочетание трех традиций

1) **разработки информационно-поисковых тезаурусов**
   (описание терминов, многословные термины, простой набор отношений)
2) **разработки лингвистических ресурсов типа WordNet**
   (связь понятия со значением, многоступенчатое построение лексико-терминологической системы, описание многозначности терминов)
3) **созданий формальных онтологий**
   (иерархическая система понятий, отношения онтологической зависимости, свойства транзитивности и наследования)

Этапы разработки: набор коллекции текстов

- **Для каждой науки (математика, физика, химия, биология, геология) – коллекции документов (от 3000 до 8000 документов, от 50 до 90 Мб)**

- **Источники коллекций - документы, доступные в Интернет:**
  - **материалы школьных уроков;**
  - **рефераты;**
  - **университетские лекции;**
  - **материалы специализированных сайтов**

Этапы разработки: автоматическое извлечение терминов из текстов

- **извлечение именных групп (2-3 слова) определенной синтаксической структуры (зависимое прилагательное и/или существительное в родительном падеже)**
- **извлечение именных групп произвольной структуры и длины на основе анализа частотных характеристик**
- **сопоставление с имеющимися ресурсами**

Отбор терминологии

- Верхние части частотных списков по каждой из наук (10 тысяч слов, 15 тысяч словосочетаний)
- Просмотр экспертами
- Отбрасывание
  - явных ошибок,
  - общеязыковых выражений,
  - составных конструкций, состоящих из терминов
  - величина результирующего списка – 32 тысячи слов и словосочетаний

Использование знаний, описанных в Общественно-политическом тезаурусе

- Ручная разметка поддеревьев
- Пересечение отобранных терминов и Общественно-политического тезауруса
- Замыкание отношений – добавление вышестоящих по таксономии

Эксперты

- **Эксперты в ПО vs. Инженеры по знаниям**
  - **дать определение**
  - **описать таксономические отношения**
  - **выделить общее для разных школ**
  - **провести ФОРМАЛЬНЫЙ АНАЛИЗ**

- **Примеры:**
  - ***горная порода, руда, минеральное образование (бывает еще и на зубах), природное минеральное образование***
  - ***национальный парк, лесопарк, парк***

- **Эксперты-лингвисты –**
  лингвистическая онтология –
  работа с текстами и значениями

Работа экспертов - 1

- **Источники**
  - **Загруженные списки («кандидаты»), надо либо перевести «кандидата» в основной список, либо удалить**
  - **Энциклопедии, словари, учебники**
  - **Интернет**

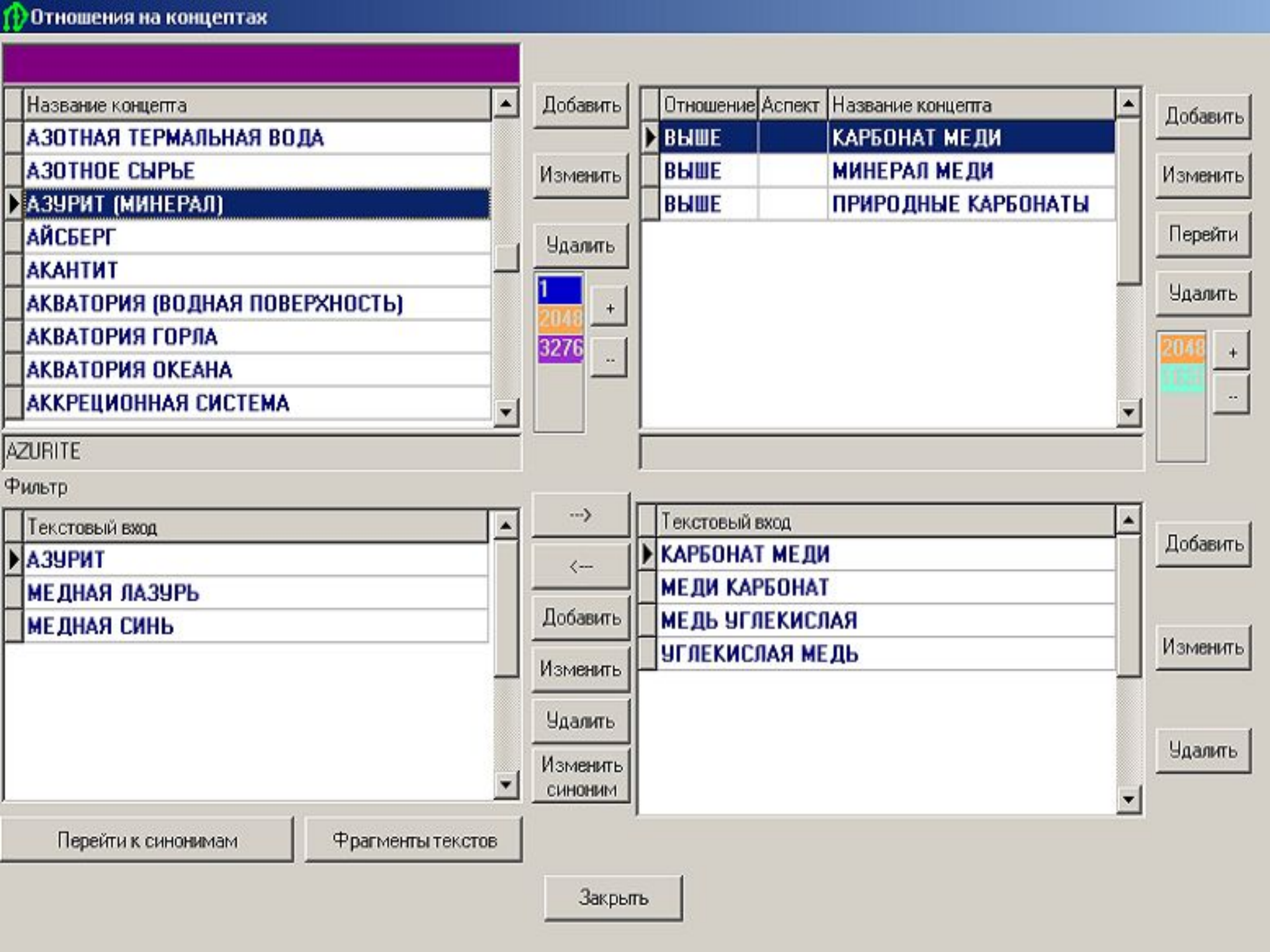
- **Операции (на основе материала источников)**
  - **Ввод нового понятия,**
  - **Описание его текстовых вариантов (макс. полно)**
  - **Таксономические отношения**
  - **Отношения зависимости понятий (на основе анализа определений, употребления в тексте**

Работа экспертов - 2

1) **Ввод нового понятия**
   - ❖ **Список «кандидатов»**
   - ❖ **Энциклопедии, книги**
2) **Поиск определения**
   - ❖ **Энциклопедии, Интернет**
   - ❖ **Анализ определения (анализ контекста употребления)**
     - **проверка определения – разные определения, старые определения**
     - **неполно выраженные, только в смысле текущего документа или в смысле подобласти**
   - **Выделение связанных понятий**
3) **Проверки**
   - **Употребляемость (Интернет, списки «кандидатов»)**
   - **Анализ лексической многозначности**

     эвтектика **(сплав vs. точка эвтектики)**

     триасс **(эпоха vs. пласт)**


Отношения на концептах
Название концепта
АЗОТНАЯ ТЕРМАЛЬНАЯ ВОДА
АЗОТНОЕ СЫРЬЕ
АЗУРИТ (МИНЕРАЛ)
АЙСБЕРГ
АКАНТИТ
АКВАТОРИЯ (ВОДНАЯ ПОВЕРХНОСТЬ)
АКВАТОРИЯ ГОРЛА
АКВАТОРИЯ ОКЕАНА
АККРЕЦИОННАЯ СИСТЕМА
Добавить
Изменить
Удалить
1
2048
3276
+
..
Отношение
Аспект
Название концепта
ВЫШЕ
КАРБОНАТ МЕДИ
ВЫШЕ
МИНЕРАЛ МЕДИ
ВЫШЕ
ПРИРОДНЫЕ КАРБОНАТЫ
Добавить
Изменить
Перейти
Удалить
2048
+
..
AZURITE
Фильтр
Текстовый вход
АЗУРИТ
МЕДНАЯ ЛАЗУРЬ
МЕДНАЯ СИНЬ
--->
<---
Добавить
Изменить
Удалить
Изменить синоним
Текстовый вход
КАРБОНАТ МЕДИ
МЕДИ КАРБОНАТ
МЕДЬ УГЛЕКИСЛАЯ
УГЛЕКИСЛАЯ МЕДЬ
Добавить
Изменить
Удалить
Перейти к синонимам
Фрагменты текстов
Закрыть

Покрытие предметной области

	понятийная структура		терминология и лексика	
	2004, X тыс.	2005,VI тыс.	2004, X тыс.	2005,VI тыс.
Всего	62,7	65,0	116,7	132,7
из них из ОПТ	24,3	24,3	74,0	74,0
«кандидаты»	56,1	43,7	106,8	88,2
Науки (без «кандидат»)	--	14,4	--	34,2
из них из ОПТ	--	4,0	--	12,0
Итого, вкл. «географию»	6,6	21,1	9,7	44,5

Уже можно использовать с существующим ПО

☐ ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ

☐ СПИСОК НАЙДЕННЫХ РУБРИК

☑ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

****	КЛЕТКА; КРОВЯНАЯ ПЛАСТИНКА; КЛЕТОЧНАЯ ТЕОРИЯ; КЛЕТОЧНОЕ ДЫХАНИЕ; РЕЦЕПТОР; ЖИЗНЕДЕЯТЕЛЬНОСТЬ КЛЕТОК; ОРГАНИЧЕСКОЕ ВОЛОКНО; ФАГОЦИТ; МЫШЕЧНОЕ ВОЛОКНО; КОСТНАЯ КЛЕТКА; ВИРУС ИММУНОДЕФИЦИТА ЧЕЛОВЕКА; ТКАНЬ; ДЕЛЕНИЕ КЛЕТОК; ДЕНДРИТ; ЦИТОЛОГИЯ; ДЕЗОКСИРИБОНУКЛЕИНОВАЯ КИСЛОТА;						
****	•	КОСТЬ; НИЖНЯЯ ЧЕЛЮСТЬ; ЖЕЛТЫЙ КОСТНЫЙ МОЗГ; ГРУДИНА; КОСТЬ ЧЕРЕПА; КРАСНЫЙ КОСТНЫЙ МОЗГ; БЕДРЕННАЯ КОСТЬ; ЗАТЫЛОК; КЛЮЧИЦА; ЧЕРЕП; ЧЕЛЮСТЬ (КОСТЬ); ВИСОК; ТАЗОВЫЕ КОСТИ; НАДКОСТНИЦА; ЛОПАТКА (КОСТЬ); РЕБРО (КОСТЬ);					
****	•	•	КРОВЬ; КРОВЬ ЧЕЛОВЕКА; СОВМЕСТИМОСТЬ КРОВИ; СВЕРТЫВАНИЕ КРОВИ; ЛИМФАТИЧЕСКИЙ СОСУД; ЛИМФАТИЧЕСКИЙ УЗЕЛ; ГАММА-ГЛОБУЛИН; ГЕМОФИЛИЯ; КРОВЕТВОРНАЯ СИСТЕМА; ГРУППА КРОВИ; СЕЛЕЗЕНКА; РЕЗУС-ФАКТОР; ТРОМБ; ЛИМФА; ПЛАЗМА КРОВИ; АНЕМИЯ;				
****	•	•	·	МЫШЦА; МЫШЦА ГОЛОВЫ; ПОПЕРЕЧНАЯ МЫШЦА; МЫШЦА КОНЕЧНОСТЕЙ; МЫШЕЧНАЯ СИСТЕМА; МИМИЧЕСКАЯ МЫШЦА; СКЕЛЕТНАЯ МЫШЦА; СИСТЕМА ОРГАНОВ; МЕЖРЕБЕРНАЯ МЫШЦА; ДЫХАТЕЛЬНАЯ МЫШЦА; ТОНУС (ХАРАКТЕРИСТИКА МЫШЦ И НЕРВНЫХ ЦЕНТРОВ); ДИАФРАГМА; ГЛАДКАЯ МЫШЦА; ФИЗИЧЕСКАЯ СИЛА; БРЮШНОЙ ПРЕСС; СУХОЖИЛИЕ;				
****	•	•	·	·	ГОЛОВНОЙ МОЗГ; ЧАСТЬ ТЕЛА; ОРГАНИЗМ ЧЕЛОВЕКА; ЦЕНТРАЛЬНАЯ НЕРВНАЯ СИСТЕМА; ГИПОФИЗ; НЕРВНАЯ ТКАНЬ; МОЗЖЕЧОК; ГОЛОВА (ЧАСТЬ ТЕЛА); НЕРВНАЯ СИСТЕМА;		
****	•	•	·	·	·	ТКАНЕВАЯ ЖИДКОСТЬ; ЖИДКАЯ СРЕДА ОРГАНИЗМА; ЖИДКОСТИ; ЖИВОТНОЕ; ВЕЩЕСТВО;	
****	·	·	·	·	·	·	ДВИЖЕНИЕ, ПЕРЕМЕЩЕНИЕ; КОЛЕБАТЕЛЬНОЕ ДВИЖЕНИЕ; СКОРОСТЬ (СТЕПЕНЬ БЫСТРОТЫ ДВИЖЕНИЯ); ЧАСТОТА (ВЕЛИЧИНА, ОТРАЖ. ЧИСЛО ПОВТОРЕНИЙ В ЕДИНИЦУ ВРЕМЕНИ; МЕМБРАНА (ДЕТАЛЬ УСТРОЙСТВА);

☐ АННОТАЦИЯ

▼ ☑ ОБРАБОТАННЫЙ ТЕКСТ

Эпителиальная ткань образует **покровы тела**, **железы**, выстилает **полости внутренних органов**. **Клетки ткани** близко прилегают друг к другу, **межклеточного вещества** мало. Создается препятствие для проникновения **микробов**, **вредных веществ**, защита лежащих под **эпителием тканей**. Смена **клеток** происходит благодаря способности к быстрому **размножению**.

Соединительная ткань. Ее особенность - сильное развитие **межклеточного вещества**. **Основные** функции **ткани** - питательная и опорная. К **соединительной ткани** относятся **кровь**, **лимфа**, **хрящевая**, **костная**, **жировая ткани**.

Кровь и **лимфа** состоят из **жидкого межклеточного вещества** и **клеток крови**. Эти **ткани** обеспечивают связь между органами, перенося **вещества** и **газы**.

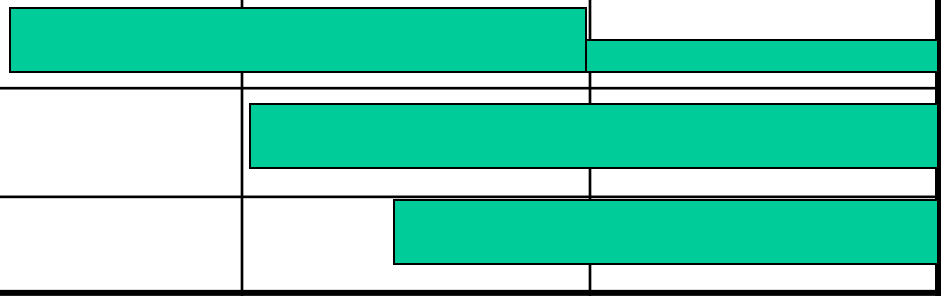
Волокнистая соединительная ткань состоит из **клеток**, связанных **межклеточным веществом** в виде

Обсуждаемые применения

- **Мониторинг**
  - **инновационно ориентированный мониторинг, установление связей между сервисами / продуктами и результатами научных исследований**
  - **поддержка экспертизы заявок и отчетов научных проектов, исключение дублирования**
- **Техническое регулирование**
  - **поддержка экспертизы технических регламентов**
  - **определение объектов техрегулирования в тексте**
  - **применимость того или иного технического регламента**
  - **определение нарушения требований технических регламентов**

Выводы

план	2005	2006	2007
покрытие			
тестирование			
приложения			



- **Представлено состояние работ разработки лингвистической онтологии большого размера для автоматической обработки текстов научно-технической тематики**

- **Ресурс будет бесплатен для некоммерческого применения**

Отношения онтологической зависимости. Формальная онтология. N.Guarino

- **может ли сущность (*C1*) существовать сама по себе, или подразумевает существование чего-либо еще (*C2*):**
- **подразумевает ли существование сущности существование какой-либо конкретной сущности (*строгая зависимость* - rigid dependence)**
  жидкость (C1) - кипение (C2),
  ***минерал – геологическое отложение*,**
- **предполагается ли существование примеров некоторого класса (generic dependence – *зависимость по классу*) некоторых сущностей,**
  землетрясение (C1) – шкала Рихтера (C2); вулкан – вулканология, газовое месторождение – газовая разведка.
- **взаимозависимые понятия:** ***катализ - катализатор***