



**РОССИЙСКИЙ НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР
ПО ЛИНГВИСТИКЕ ИМ. И.А.БОДУЭНА ДЕ КУРТЕНЭ**



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



Казанский
государственный
университет
им. В.И.Ульянова- Ленина

**Автоматизированное индексирование описаний
музейных предметов на
базе русскоязычной версии
Тезауруса по архитектуре и искусству
(Тезауруса ААТ)**

Добров Б.В. , Лукашевич Н.В., Соловьев В.Д.

louk@mail.cir.ru

Доступ к цифровым ресурсам по культурному наследию

- **Объекты нетекстовой природы**
 - Текстовые описания
 - Поиск по изображениям
- **1-5% музейных экспонатов выставлено в экспозициях**
- **Лингвистические ресурсы для концептуального индексирования**
 - Тезаурусы
 - Онтологии

Тезаурус по архитектуре и искусству (тезаурус ААТ)

- **Объем: 30 тысяч дескрипторов; 130 тысяч англоязычных терминов**
- **Терминология по искусству, архитектуре, материальной культуре, архивным материалам с античности до наших дней.**
- **Наиболее полное покрытие: искусство Западной Европы и Америки**
- **Специфика искусства народов России представлена недостаточно**
- **Но перечислено множество общезначимых сущностей: материалов, объектов материальной культуры и искусства**

Адаптация Тезауруса ААТ для описания культуры народов России

- **Перевод на русский язык**
- **Дополнение русскоязычными синонимами**
 - **Общезначимый русский язык (ручка – рукоятка – черенок)**
 - **Музейная терминология**
- **Дополнение специальной терминологией – отражение специфики культуры России**
- **Современные тенденции в развитии ресурсов:**
 - **Сбор текстовых коллекций (корпусов – каталоги, описания музейных предметов)**
 - **Автоматизированное извлечение терминов по текстам**

Информационная система «Культурное наследие РОССИИ»

- **Научно-образовательный центр по лингвистике при Казанском государственном университете**
- **НИВЦ МГУ- опыт:**
 - **Автоматизированная разработка терминологических ресурсов по текстовым коллекциям**
 - **Создание тезаурусов и онтологий для автоматического концептуального индексирования**
 - **Тезаурус русского языка РуТез – 49 тысяч понятий, 135 тысяч русскоязычных слов, выражений, терминов**
 - **Разработка информационных систем на основе технологий концептуального поиска**

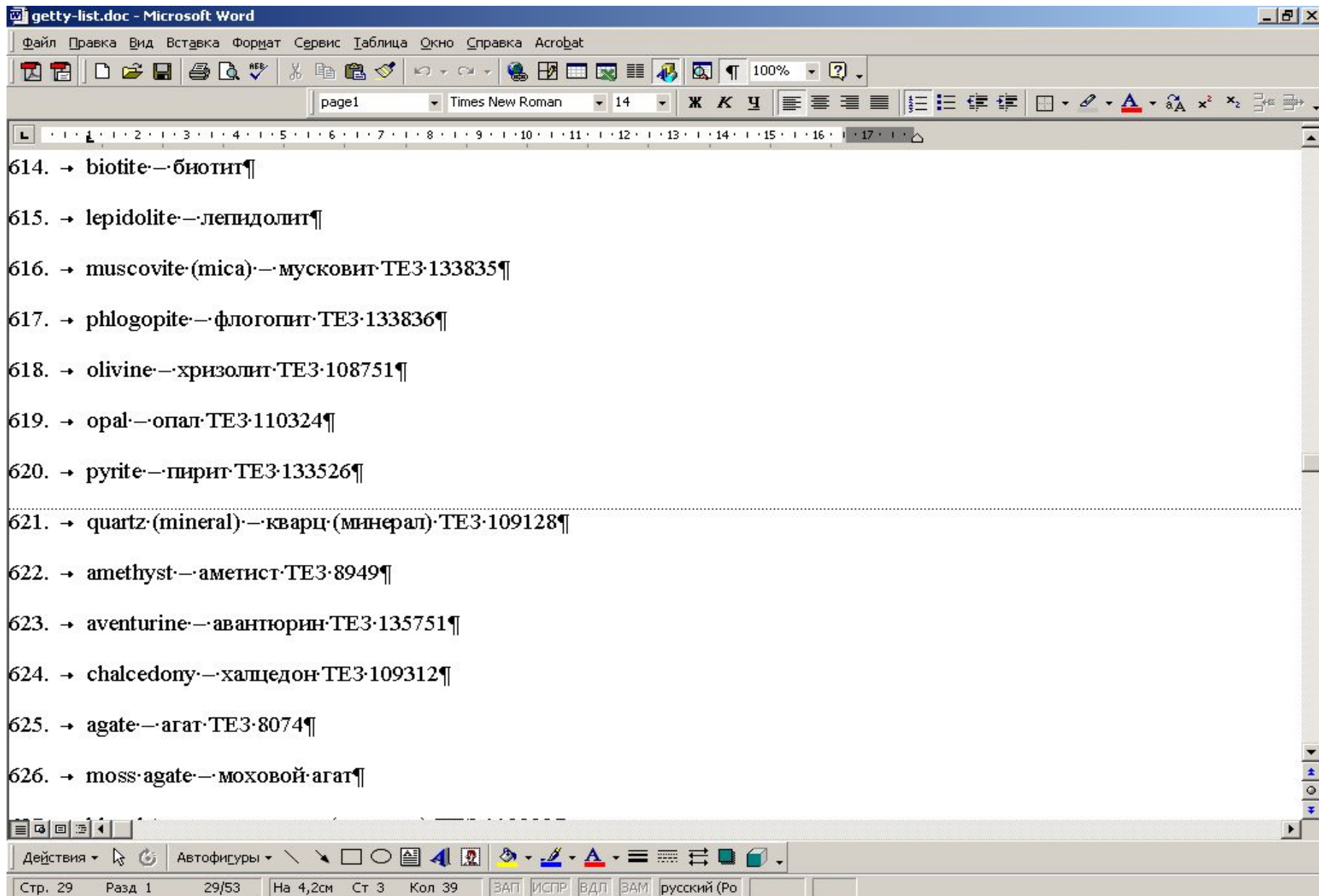
Система автоматизированного индексирования на базе тезауруса ААТ

- **Получена лицензия от фонда Гетти на некоммерческое использование тезауруса ААТ**
- **Перевод фасетов ААТ: Материалы и объекты**
- **Переведено 10 тысяч дескрипторов**
- **Ссылка на понятие тезауруса РуТез, если есть – известные общезначимые русскоязычные синонимы, дополнительные отношения**
- **Экспериментальная загрузка двуязычного ресурса в тезаурусную оболочку: исходный дескриптор – англоязычные синонимы, русскоязычный дескриптор, дополнение синонимами из Тезауруса РуТез**
- **Экспериментальная обработка реальной коллекции описаний музейных предметов**

Этапы работы системы автоматизированного индексирования

- Графематический анализ текста – разбиение текста на значимые элементы: слова, знаки препинания, числа и т.п.
- Морфологический анализ текста – приведение слов текста к словарной форме
- Терминологический анализ текста – сопоставление слов текста с терминами тезауруса
- Разрешение неоднозначности – *ручка: ручка чашки, перьевая ручка*
- Результат: индекс по дескрипторам тезауруса – концептуальный индекс – не зависит от исходного языка документа

Фрагмент файла перевода фасета «Материалы»



The image shows a screenshot of a Microsoft Word document titled "getty-list.doc". The document contains a list of mineral names in English and their Russian translations, numbered 614 to 626. The interface includes a menu bar, a toolbar, and a status bar at the bottom.

614. → biotite → биотит¶

615. → lepidolite → лепидолит¶

616. → muscovite (mica) → мусковит ТЕЗ-133835¶

617. → phlogopite → флогопит ТЕЗ-133836¶

618. → olivine → хризолит ТЕЗ-108751¶

619. → opal → опал ТЕЗ-110324¶

620. → pyrite → пирит ТЕЗ-133526¶

621. → quartz (mineral) → кварц (минерал) ТЕЗ-109128¶

622. → amethyst → аметист ТЕЗ-8949¶

623. → aventurine → авантюрин ТЕЗ-135751¶

624. → chalcedony → халцедон ТЕЗ-109312¶

625. → agate → агат ТЕЗ-8074¶

626. → moss agate → моховой агат¶

Стр. 29 Разд 1 29/53 На 4,2см Ст 3 Кол 39 ЗАП ИСПР ВДЛ ЗАМ русский (Ро

Экран программной оболочки ведения тезауруса

Отношения на концептах

КВАР

Название концепта
произведения изобразительного искусства
произведения изобразительного искусства
произведения изобразительного искусства
произведения изобразительного искусства
произведения изобразительного искусства
произведения искусства

visual works by form

Фильтр

Текстовый вход

Добавить

Изменить

Удалить

2048 + -

Отношения	Название концепта
НИЖЕ	диорамы
НИЖЕ	диптихи
НИЖЕ	образы предков у Африка
НИЖЕ	медали
НИЖЕ	полиптихи
НИЖЕ	триптихи

dioramas

--->

<---

Добавить

Изменить

Удалить

Изменить синоним

Текстовый вход
ДИОРАМА
ДИОРАМНЫЙ

Добавить

Изменить

Удалить

Перейти к синонимам

Фрагменты текстов

Закреть

Примеры из коллекции Казанского этнографического музея

- **Кукла из бумаги. Лицевая сторона обтянута шёлком**
- **Обезьяна, голубые глаза из бисера, покрыта кожей с волосяным покровом.**
- **Куша. Тело из пестряди. Платье из иранского ситца, с поясом.**
- **Кукла; юбка непропорционально длинная, красного цвета. Голова покрыта платком из красного ситца.**
- **Кукла. Платье из коричневой ткани. Волосы из пакли, заплетены в косу.**
- **Кукла тряпичная. Сарафан из старой ткани розового цвета. Фартук и кофта из красного ситца с беленькими цветочками. На голове розово-белый платок.**

Экранная форма ввода описания предмета



USER: **FREE**
Доступ:
FREE
Имя: Пароль:
[] [] []
[Регистрация](#)
[Забыли пароль?](#)

[ПОИСК](#)

Автоматическое описание через дескрипторы Тезауруса AAT (The Art & Architecture Thesaurus)



Введите
текст:

Кукла; юбка непропорционально длинная, красного цвета. Голова покрыта платком из красного ситца.

Кукла. Платье из коричневой ткани. Волосы из пакли, заплетены в косу.

Кукла тряпичная. Сарафан из старой ткани розового цвета. фартук и кофта из красного ситца с беленькими цветочками. На голове розово-белый платок.

Кукла в красном ситцевым сарафане, чёрном кафтане и чёрной повязке.

Кукла малая, связанная из соломы. Голова обнута коричневой и белой материей, сшитой в районе затылка. Передник из белой материи.

Молоток деревянный с деревянный наконечником.

Загрузить URL:

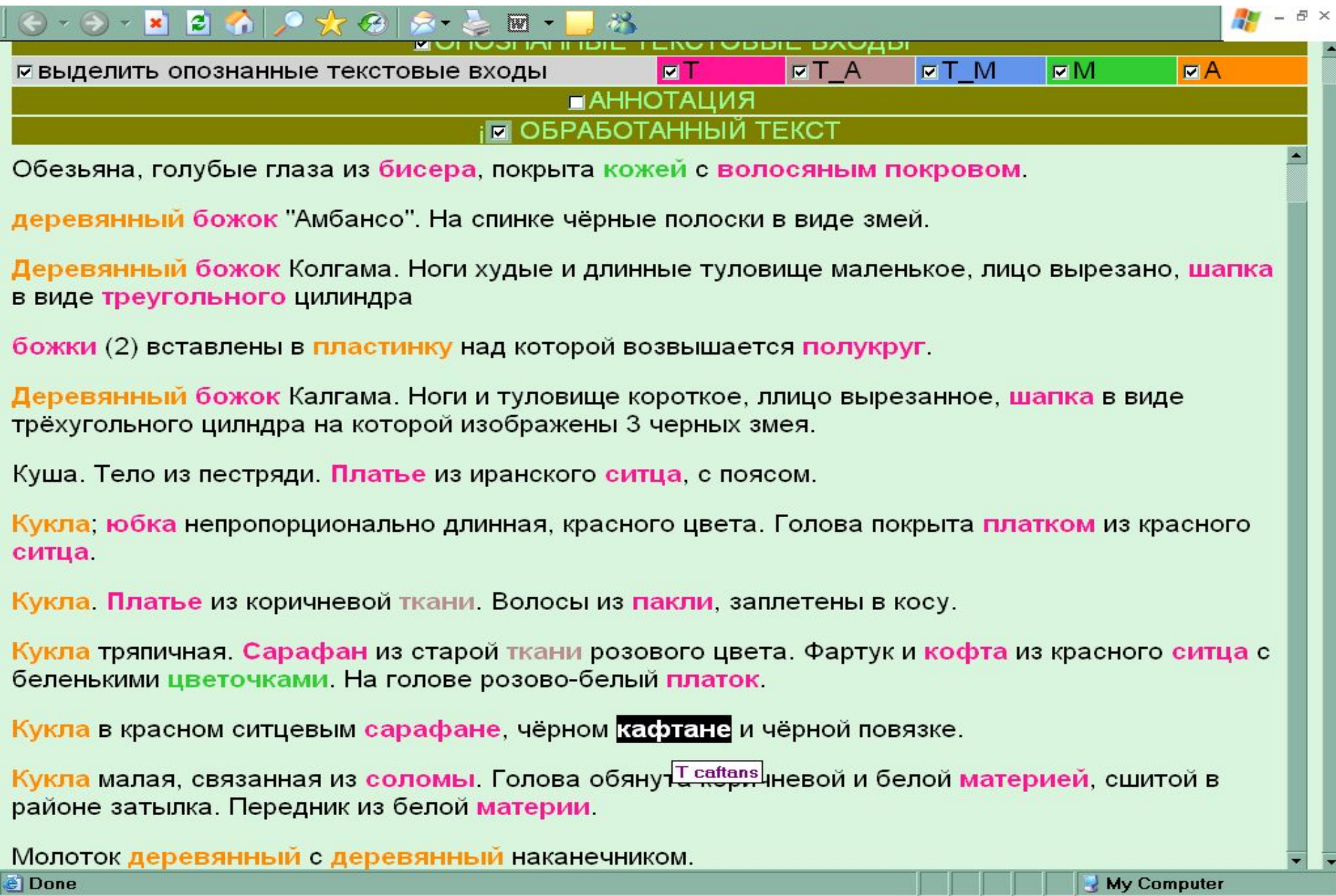
Загрузить файл:

"давить" ссылки

Пример работы терминологического анализа

hair	ВОЛОСЯНОЙ	ПОКРОВ
wood	ДЕРЕВЯННЫЙ	
cult images	БОЖОК	
wood	ДЕРЕВЯННЫЙ	
cult images	БОЖОК	
headdresses	ШАПКА	
triangles	ТРЕУГОЛЬНЫЙ	
cult images	БОЖОК	
phonograph records	ПЛАСТИНКА	
semicircles	ПОЛУКРУГ	
wood	ДЕРЕВЯННЫЙ	
cult images	БОЖОК	
headdresses	ШАПКА	
dresses	ПЛАТЬЕ	
chintz	СИТЕЦ	
puppets	КУКЛА	
skirts	ЮБКА	
kerchiefs	ПЛАТОК	
chintz	СИТЕЦ	

Результаты автоматической обработки



☐ выделить опознанные текстовые входы Т Т_А Т_М М А

☐ АННОТАЦИЯ

☐ ОБРАБОТАННЫЙ ТЕКСТ

Обезьяна, голубые глаза из **бисера**, покрыта **кожей** с **волосяным покровом**.

деревянный божок "Амбансо". На спинке чёрные полосы в виде змей.

Деревянный божок Колгама. Ноги худые и длинные туловище маленькое, лицо вырезано, **шапка** в виде **треугольного** цилиндра

божки (2) вставлены в **пластинку** над которой возвышается **полукруг**.

Деревянный божок Калгама. Ноги и туловище короткое, лицо вырезанное, **шапка** в виде трёхугольного цилиндра на которой изображены 3 черных змея.

Куша. Тело из пестряди. **Платье** из иранского **ситца**, с поясом.

Кукла; **юбка** непропорционально длинная, красного цвета. Голова покрыта **платком** из красного **ситца**.

Кукла. **Платье** из коричневой **ткани**. Волосы из **пакли**, заплетены в косу.

Кукла тряпичная. **Сарафан** из старой **ткани** розового цвета. Фартук и **кофта** из красного **ситца** с беленькими **цветочками**. На голове розово-белый **платок**.

Кукла в красном ситцевым **сарафане**, чёрном **кафтани** и чёрной повязке.

Кукла малая, связанная из **соломы**. Голова обнута **кафтани** и белой **материей**, сшитой в районе затылка. Передник из белой **материи**.

Молоток **деревянный** с **деревянный** наконечником.

Done My Computer

Заключение

- **Тезаурус по архитектуре и искусству – важный источник общезначимых терминов в сфере материальной культуры**
- **Для адаптации тезауруса ААТ для описания объектов материальной культуры России: необходим не только перевод на русский язык, но и пополнение русскоязычными синонимами и специфическими терминами**
- **Существенной базой для автоматизированного пополнения могут служить электронные коллекции текстов музеев: каталоги, описания**
- **Сотрудничество: наш опыт работы автоматической обработки текстов, разработки тезаурусов + музеи: коллекции, терминология=> двуязычный тезаурус по архитектуре и искусству**