Системы информационного обеспечения для проведения распределенных научных исследований

План доклада

- Задачи систем информационного обеспечения распределенных научных исследований
- Тематический поиск информации
- Технологии построения специализированных информационных систем для проведения распределенных научных исследований

Задачи информационного обеспечения распределенных научных исследований

- Сбор информации из внешних и внутренних источников о предмете исследования
- Автоматизация процесса обмена файлами данных между участниками
- Автоматизация процесса распределенной обработки информации

План доклада

- Задачи систем информационного обеспечения распределенных научных исследований
- Тематический поиск информации
- Технологии построения специализированных информационных систем для проведения распределенных научных исследований

Задачи автоматизированной системы тематического анализа информации

- извлечение данных из текстовых коллекций на естественных языках (русский, английский)
- автоматическое обучение системы и отчуждение персональных знаний у эксперта (знаний об объектах и структуре используемого языка, онтологии)
- автоматический тематический анализ информации и выделение фактов
- перманентный автоматический подбор материалов по запросам пользователей
- сохранение версий интересующей пользователя информации
- Минимальная «ручная» настройка

Возможности системы

- Разовый поиск информации
- Непрерывный сбор информации
- Группировка и предварительная обработка информации для последующего анализа
- Классификация информации по древовидным классификаторам
- Контекстный поиск
- Персонифицированный поиск

Особенности системы

- Автоматическая рубрикация текстов с использованием неограниченного количества рубрикаторов
- Наличие персональных рубрикаторов и архивов
- Автоматическое ранжирование с учетом предпочтений пользователя
- Сервисно-ориентированная архитектура (SOAP)
- Открытый код

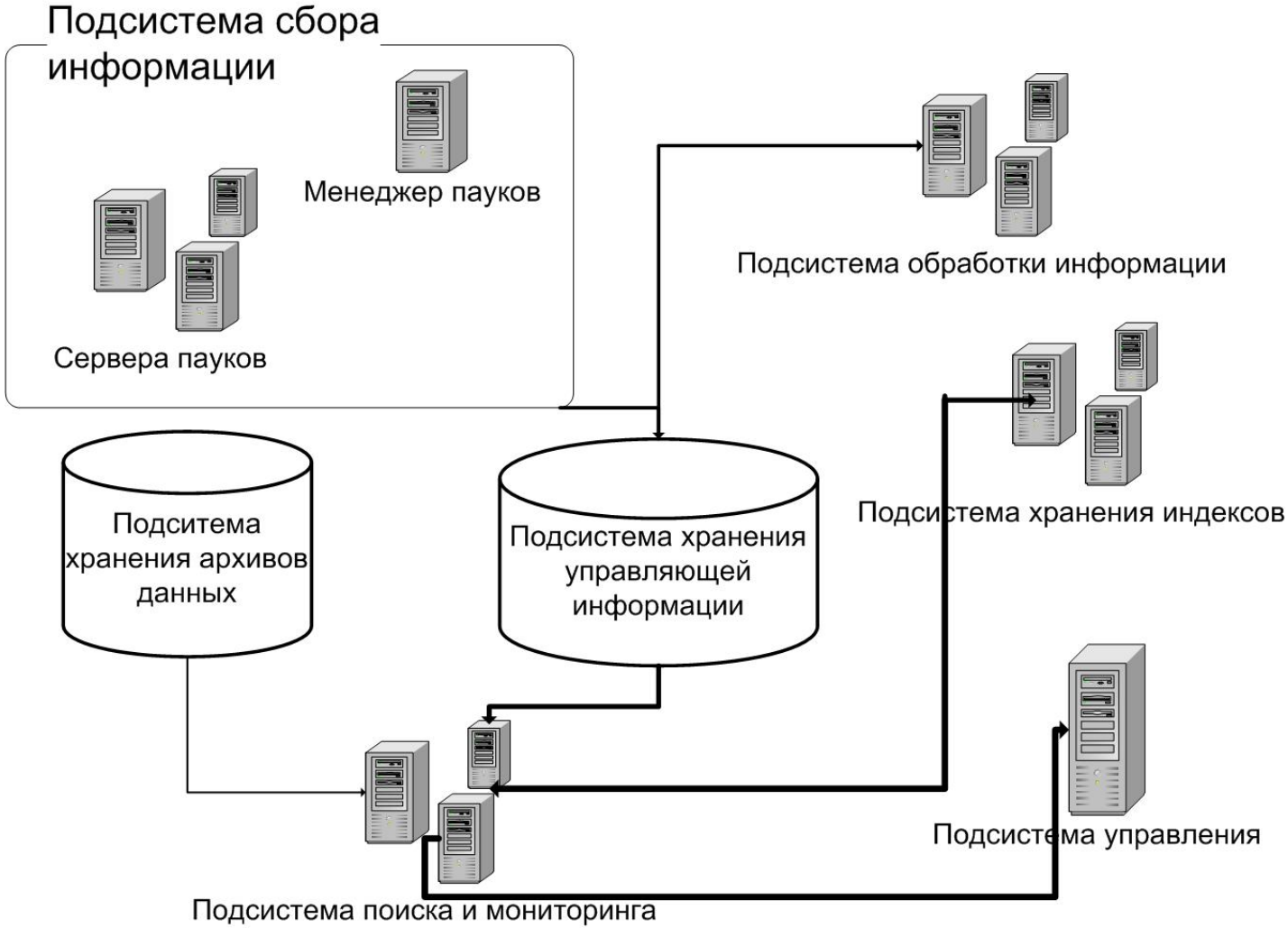
Перспективные направления

- Выявление новых тематических направлений
- Выделение дат, имен, географических объектов, источников и других параметров
- Выделение границ ресурсов
- Выделение фактов и определение их достоверности

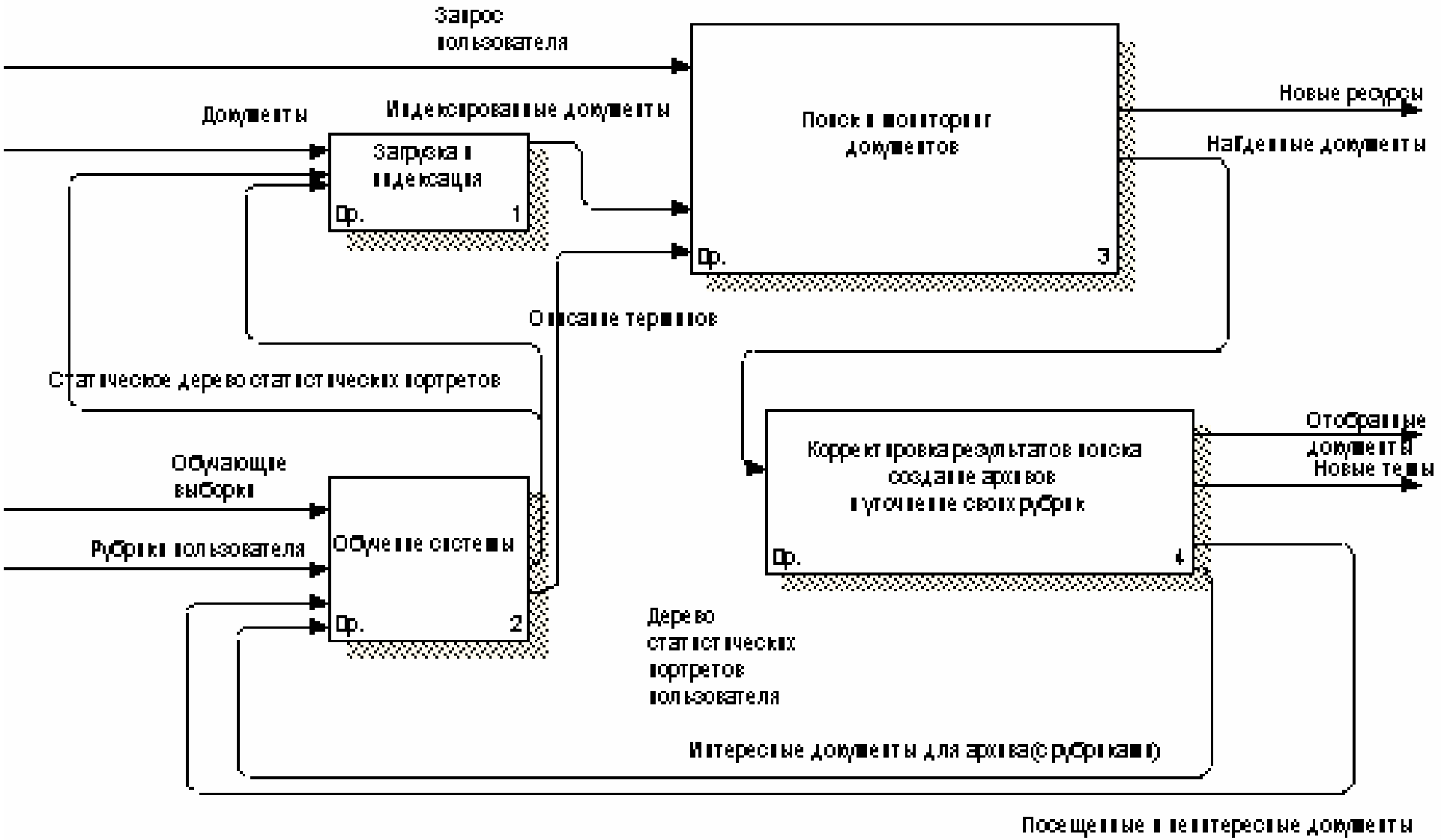
Возможное применение

- Поиск в интернет
- Поиск по новостям, форумам, периодическим изданиям
- Поиск по специализированным хранилищам
- Внедрение в организациях, требующих сертификации ПО в области информационной безопасности

Структура системы



Общая схема


Запрос пользователя
Документы
Индексированные документы
Загрузка и индексация
Др.
1
Поиск и мониторинг документов
Др.
3
Новые ресурсы
Найденные документы
Описание терминов
Статическое дерево статистических портретов
Корректировка результатов поиска создание архивов и уточнение своих рубрик
Др.
4
Отобранные документы
Новые темы
Обучающие выборки
Рубрики пользователя
Обучение системы
Др.
2
Дерево статистических портретов пользователя
Интересные документы для архива(с рубриками)
Посещенные и неинтересные документы

Основные элементы

- Распределенное хранилище
- Модуль вноса и индексации(морфологический анализ, рубрикация, аннотирование)
- Модуль обучения
- Рубрики пользователя
- Система обратной связи
- Система поиска
- Мониторинг
- Модуль тематического анализа
- Web-интерфейс
- Модуль персонификации

Загрузка

- Морфологический разбор текстов и загрузка пословного индекса
- Составление и загрузка аннотаций
- Выделение и загрузка ссылок документа
- Выделение и загрузка терминов документа

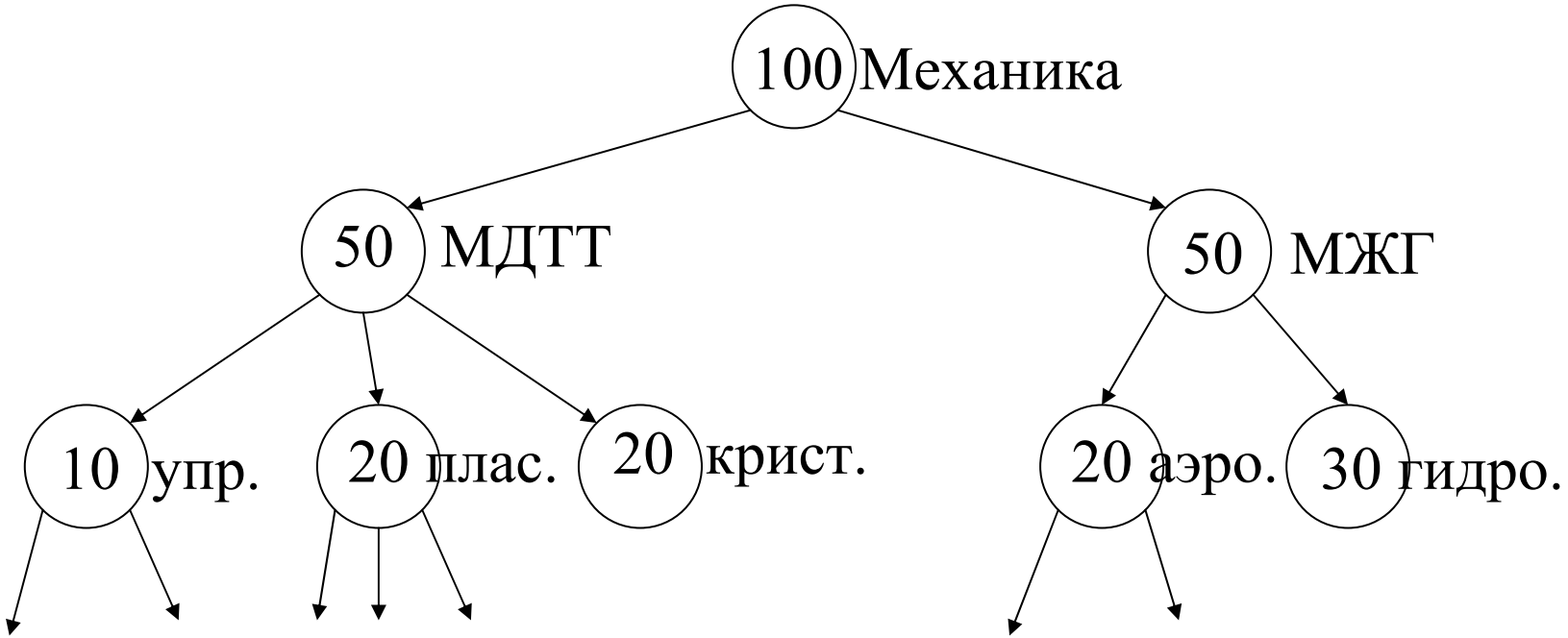
Выделение терминов

- Вычисляются все пары, входящие в обучающие выборки
- Определяются устойчивые пары
- Определяются пары, характеризующие рубрики

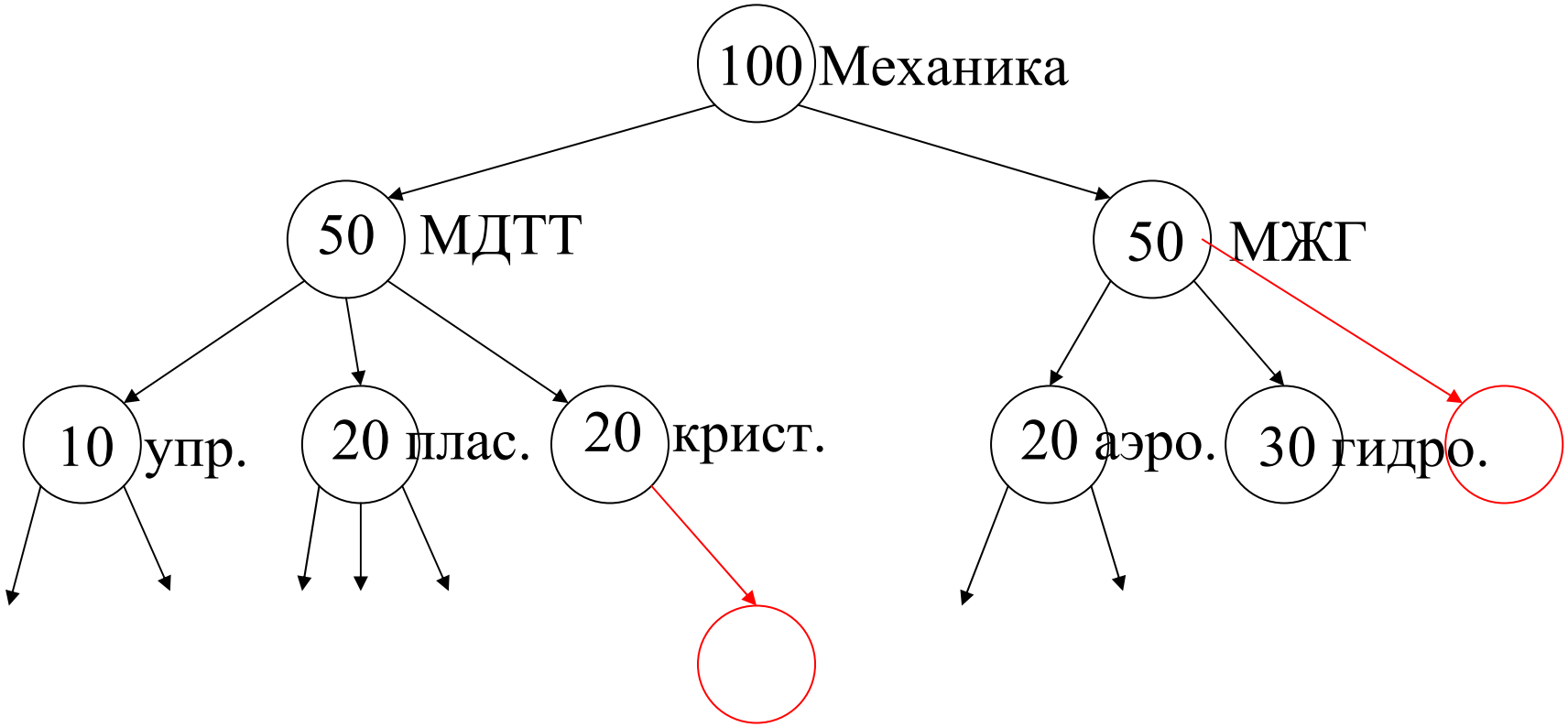
Рубрикация(1)

- Обучение
  - Составление классификатора
  - Подбор примеров для обучения
  - Построение статистических портретов рубрик
  - Построение дерева статистических портретов рубрик
- Автоматическая рубрикация входного потока текстов
- Дообучение системы

Рубрикация(2)



Рубрики пользователя



- Улучшают результаты поиска
- Позволяют использовать индекс

Язык запросов

- Запрос может содержать выражения вида
  X<w1,n1> and X<w2,n2> or X<w3,n3>
  X – тип аргумента (леммы, термины, рубрики, даты, параметры, ресурсы или ключи)
  w – символьное или числовое значение аргумента
  n – значимость при поиске

L<нефть> AND (RID<786, 0.5> OR RID<966>)

Мониторинг

- Запоминает запрос пользователя
- Регулярно обходит интересующие пользователя ресурсы
- Создает подборки интересующих пользователя документов

Система поиска

- Разбор запроса и построение дерева
- Считывание всех файлов индексов, упоминающихся в запросе
- Вычисление запроса для каждого документа
- Вычисление ранга документа
- Сортировка с усечением (Top100)
- Уточнение рангов по пользовательским рубрикам

Web-интерфейс

- Поиск
- Управление мониторингом
- Управление поиском новых ресурсов
- Конфигурация рубрикатора пользователя
- Управление архивом пользователя
- Обратная связь (посещенные и неинтересные документы)

Паук

- Менеджер пауков и группы пауков с разными приоритетами
- Правило включает маску URL, разрешение или запрещение сканирования, принадлежность рубрикам, начальный интервал переиндексации, приоритет, глубина сканирования, общее количество
- Время следующего сканирования определяется по приоритету и времени последнего изменения
- Возможно подключать пауки других производителей

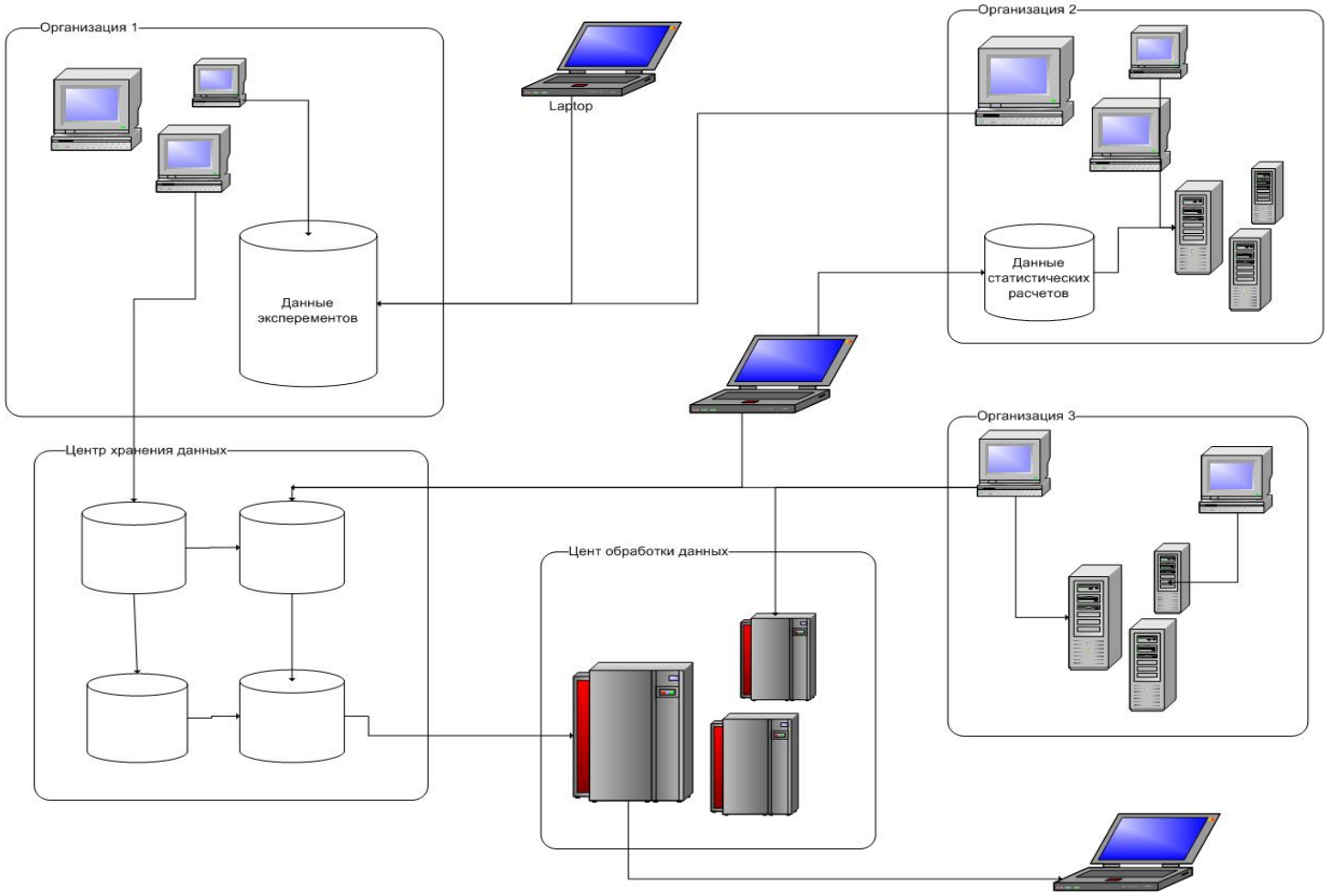
План доклада

- Задачи систем информационного обеспечения распределенных научных исследований
- Тематический поиск информации
- Технологии построения специализированных информационных систем для проведения распределенных научных исследований

Особенности строго структурированных данных

- Строгость описания
- Автоматический контроль целостности и корректности
- Простота и высокая скорость доступа при автоматической обработке
- ----------
- Невозможность создания «универсальных» структур данных
- Необходимость создания специальных информационных систем под каждую конкретную задачу

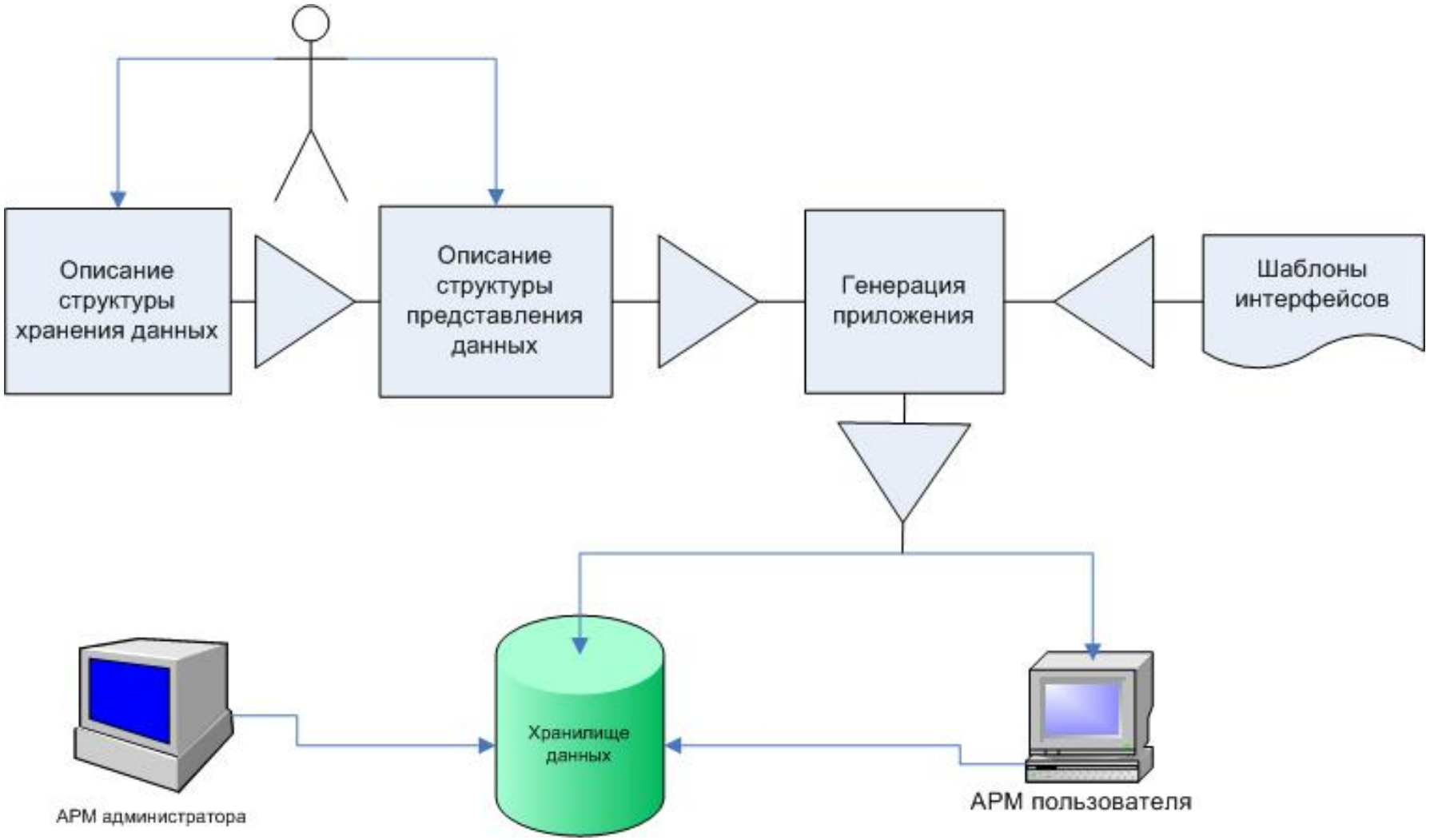
Использование справочных информационных систем



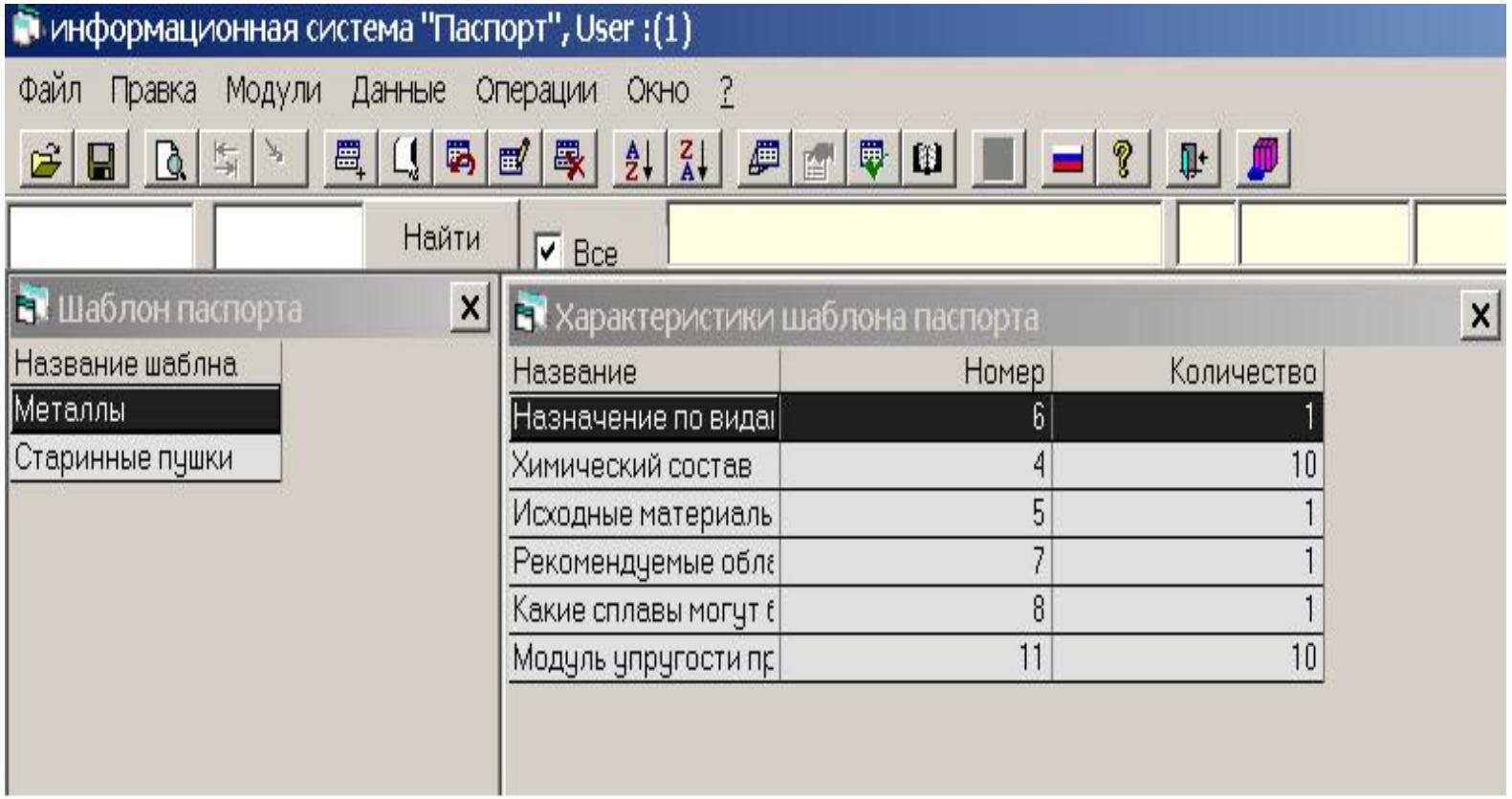
Требование к информационным системам

- Распределенный доступ к данным
- Разделение привилегий для доступа к данным
- Удобный интерфейс (поиск, сортировка, фильтрация, создание отчетов, экспорт данных)
- Быстрое создание и модификация структуры пользовательских данных

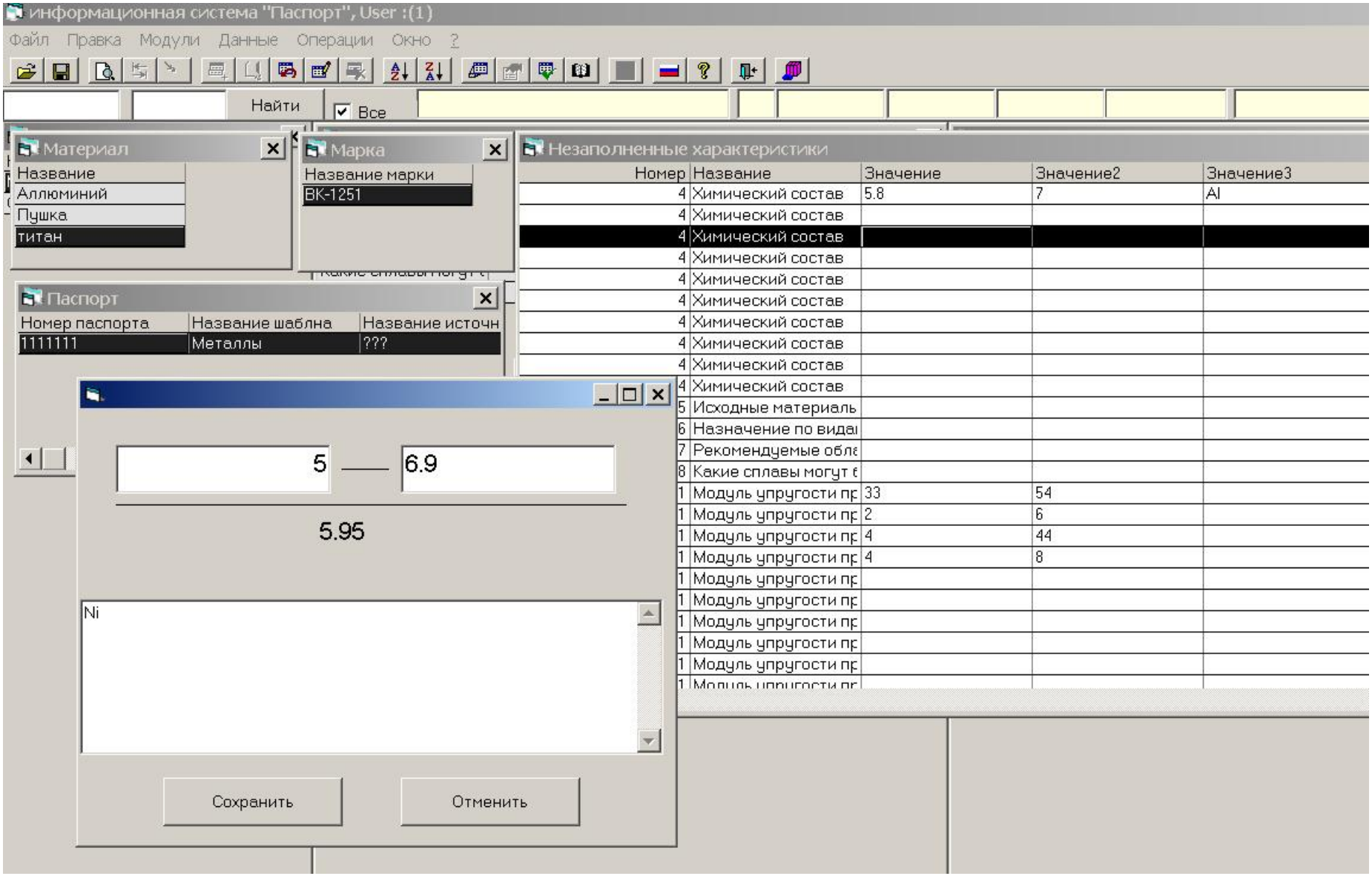
Технология быстрого создания справочных информационных систем



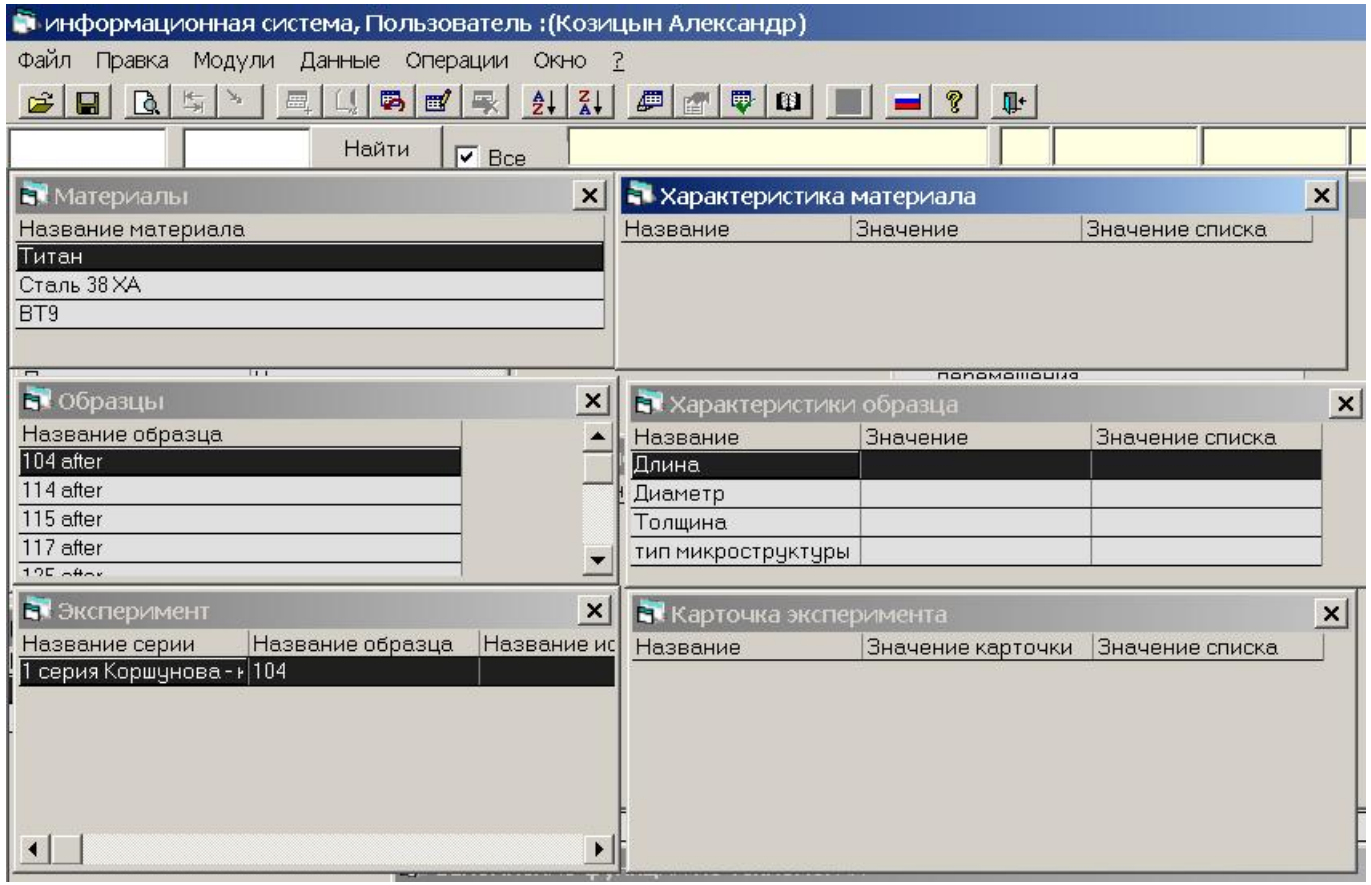
Паспорта материалов(1)



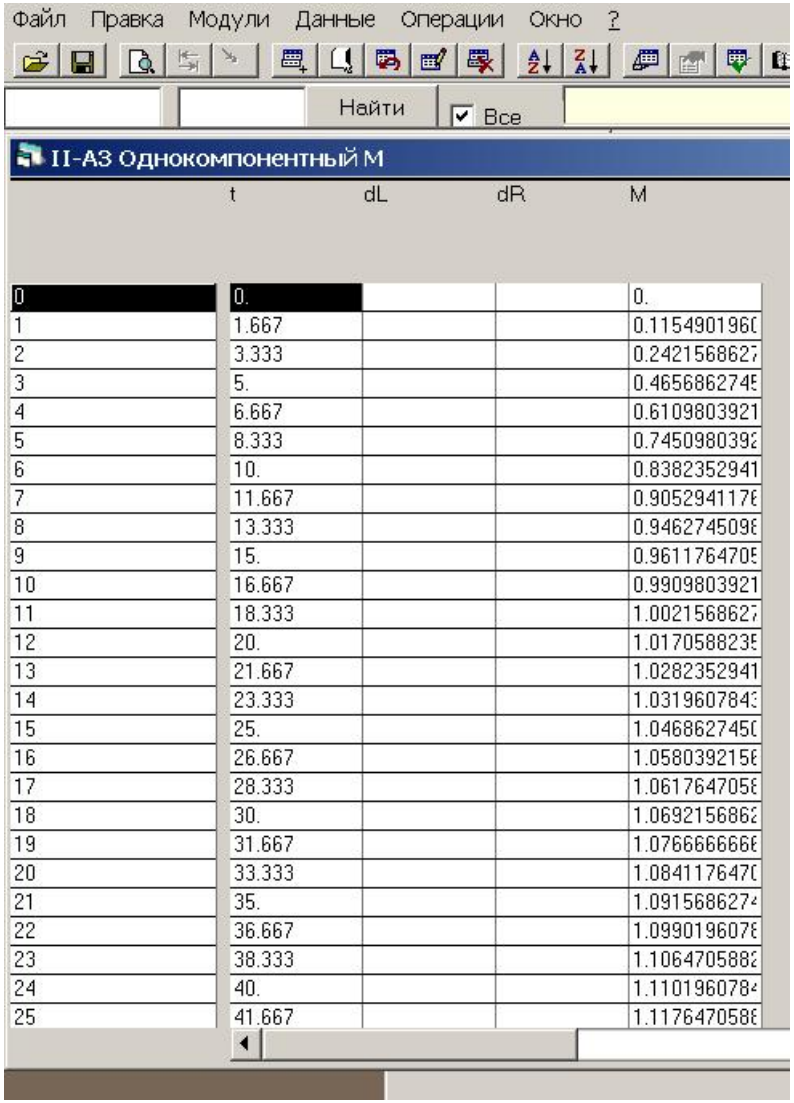
Паспорта материалов(2)



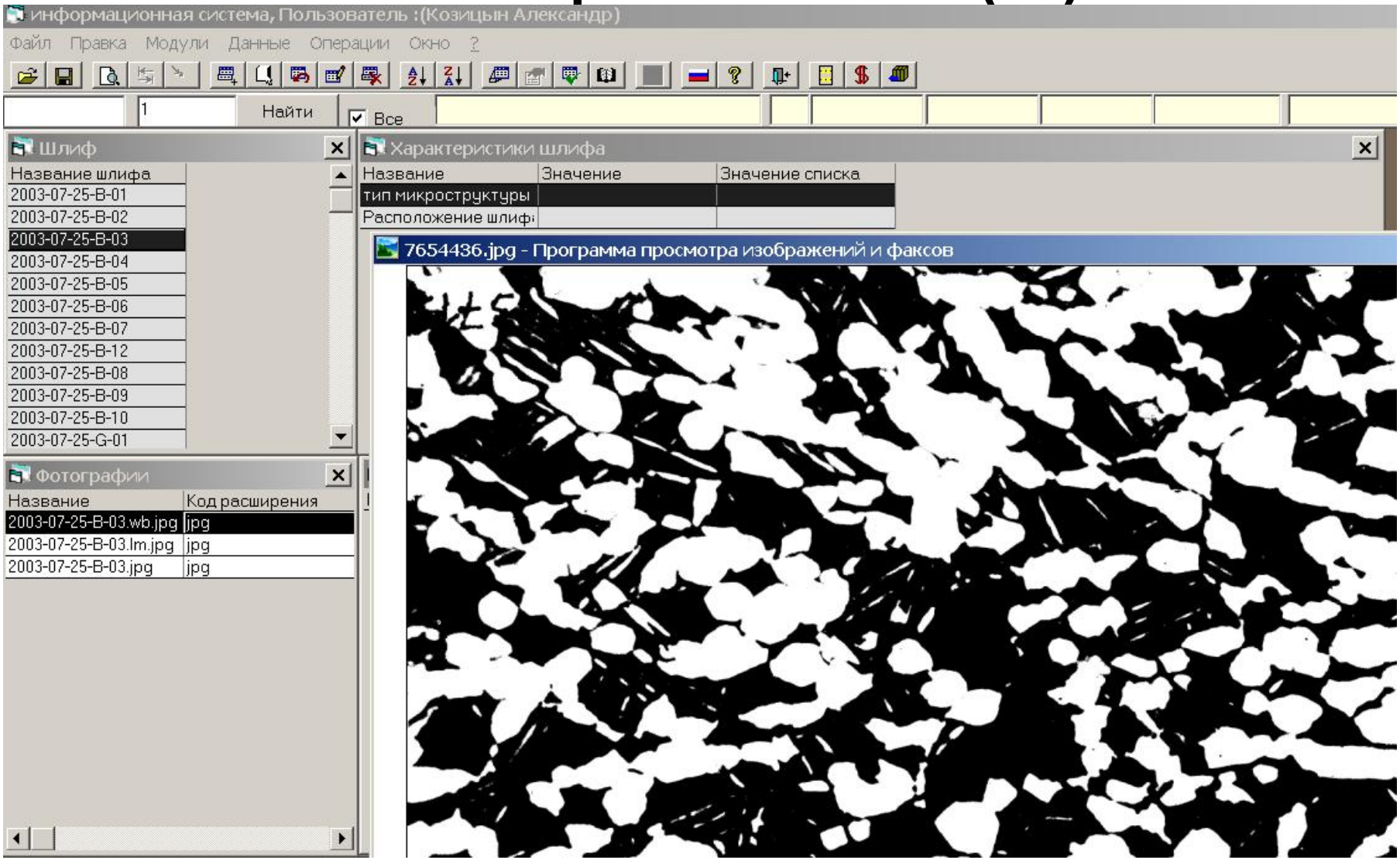
Система сопровождения экспериментов(1)



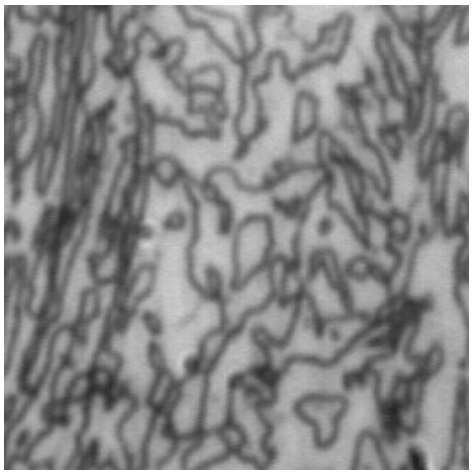
Система сопровождения экспериментов(2)

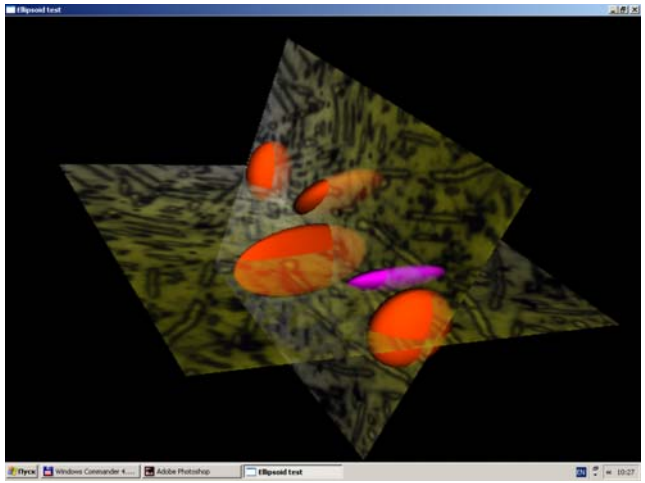


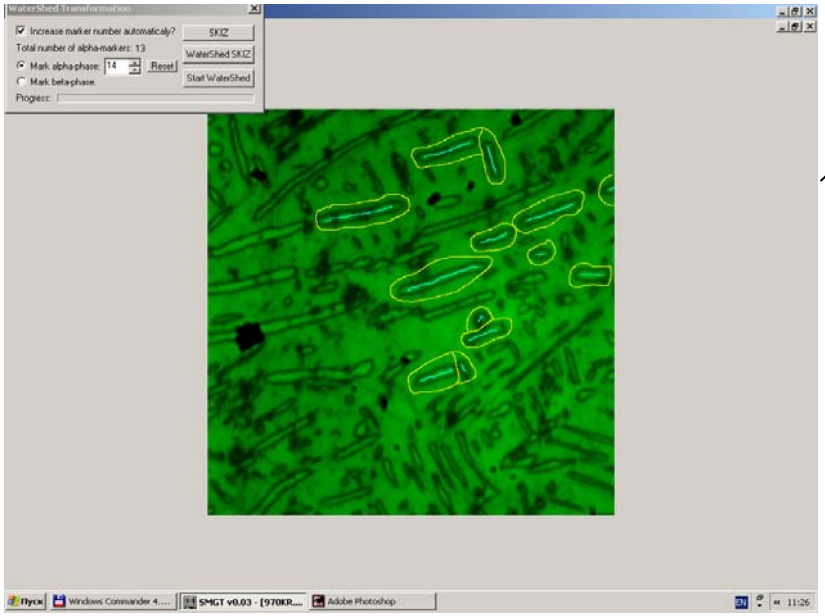
Система сопровождения экспериментов(3)



Автоматическая обработка данных

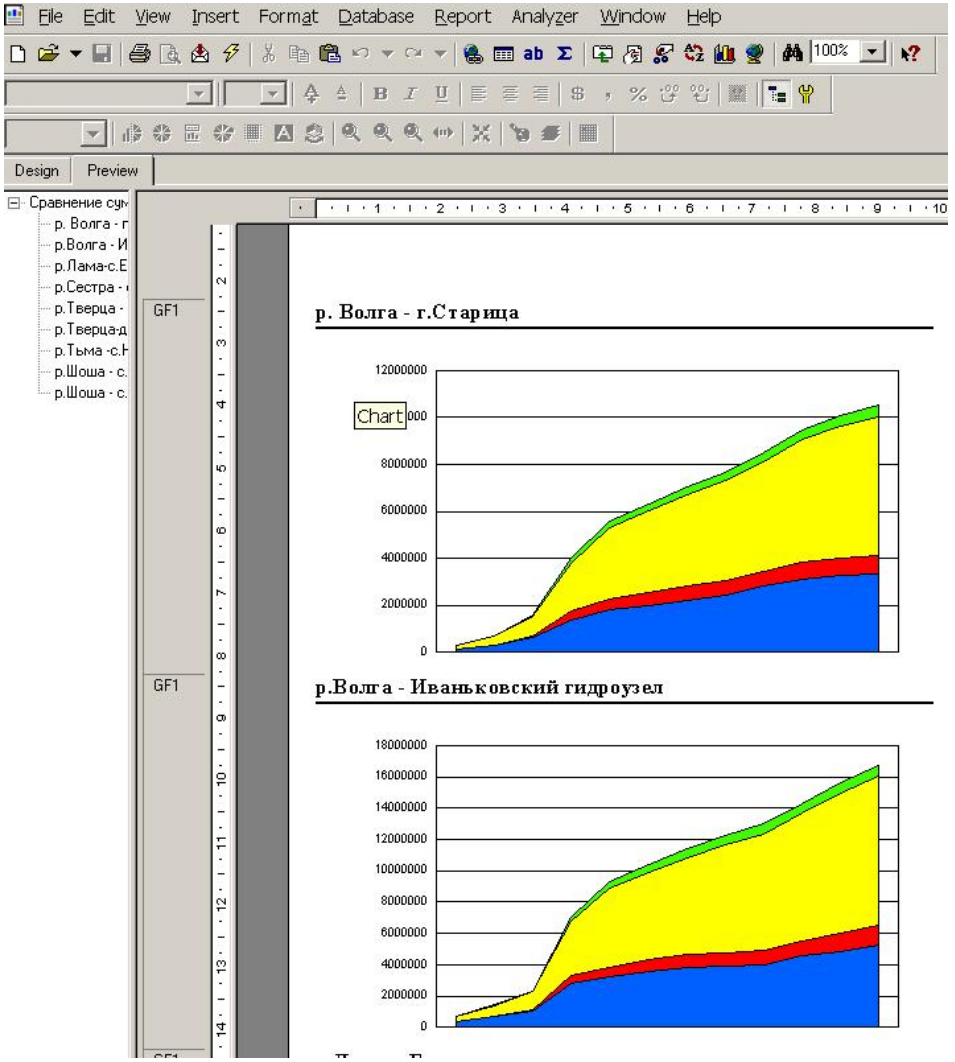




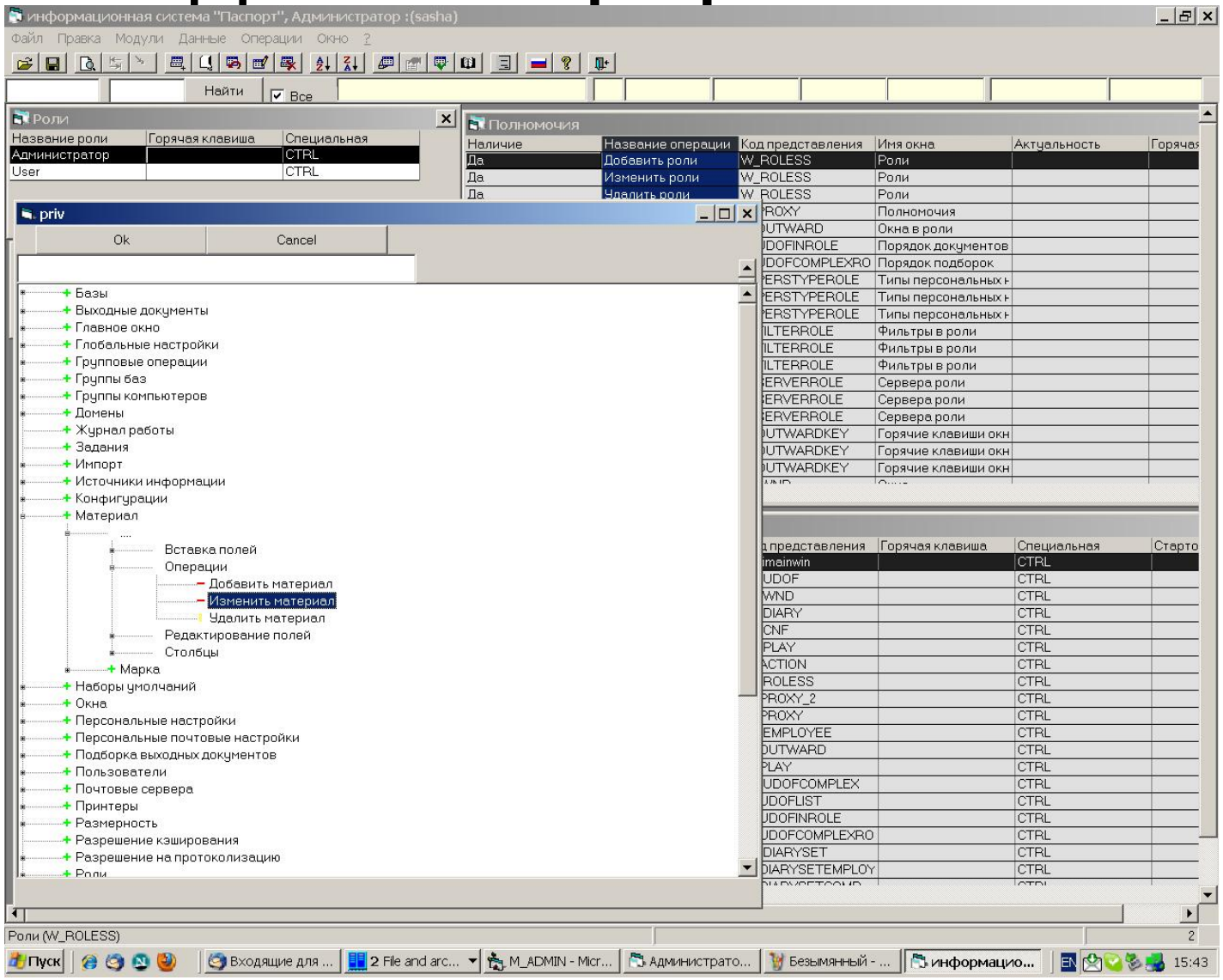


Аналитические отчеты

		р.Волга - Иваньковский гидроузел	Total
937	1	0.00	0.00
	2	0.00	0.00
	3	0.00	0.00
	4	670 000.00	670 000.00
	5	231 000.00	231 000.00
	6	74 000.00	74 000.00
	7	55 000.00	55 000.00
	8	141 000.00	141 000.00
	9	132 000.00	132 000.00
	10	95 000.00	95 000.00
	11	67 000.00	67 000.00



Подсис тема администрирования



Особенности технологии

- Простота разработки
- Визуальное представление структур данных
- Простота модификации
- Наличие шаблонов интерфейсов
- Наличие готового модуля администрирования
- ----------
- Создание только справочных систем

Спасибо за внимание