Закрытое акционерное общество «Ай-Теко»

«УТВЕРЖДАЮ»

Директор департамента систем
управления знаниями
ЗАО «Ай-Теко»

_____С.Л. Киселев

« ___» _____ 2016 г.

СИСТЕМА УПРАВЛЕНИЯ ФАКТОГРАФИЧЕСКОЙ ИНФОРМАЦИЕЙ «X-FILES»

Версия 2.8.1

Подсистема управления процессами аналитической обработки

РУКОВОДСТВО СИСТЕМНОГО ПРОГРАММИСТА

ДШСК.50 8120 9.002-02 32 02

Листов 40

Руководитель проекта

С.М.Коровенков
« ___» _____ 2016 г.

Технический писатель

Л.В. Федонина
« ___» _____ 2016 г.

Аннотация

В настоящем документе содержатся общие сведения о структуре подсистемы управления задачами извлечения знаний (далее подсистема), рассмотрены вопросы установки и настройки компонентов, входящих в ее состав.

Примечание. В связи с постоянным развитием подсистемы элементы интерфейса и значения ее фактических параметров могут отличаться от документированных.

Содержание

Определения и сокращения

AKDF	Analytical Courier Document Format. Универсальный формат представления входных документов для подсистемы ввода. Представляет собой файл архива, состоящий из трех файлов: – Карточка полей документа в формате XML; – Текст документа (TXT); – Оригинал документа. Подробнее формат описан в Руководстве пользователя.
Apache Lucene	программа обработки текста (индексирование, поиск, …)
Semantic Self Defined Sentencies, SSDS	семантически самоопределенные предложения, в которых разрешены все ссылки, нормализована синтаксическая структура и лексика.
Анализ мнений (высказываний)	выделение из текста высказываний субъекта об объекте. В качестве объекта может выступать сущность или тема.
Аннотация	извлечение из текста документа, содержащее доминантные ключевые темы. Может быть выполнена в виде snippets или важных предложений.
БД	база данных.
Дайджест	расположенные в естественном порядке предложения документа, содержащие объект мониторинга, либо референтную ссылку на него.
Досье	документ, состоящий из разделов-атрибутов, каждый из которых содержит исторически упорядоченные факты для выбранного объекта.
Задача	предписание по управлению автоматом, с установленными регламентом и параметрами выполнения. Замечание. Описание задачи содержит сведения о том, какие документы, как и когда будут обрабатываться программой-автоматом. В задаче указываются: – информационный фонд документов, – набор атрибутов для фильтрации релевантных документов (для задач системы "X-Files"), – множество целевых объектов (для задач системы "X-Files").
Информационный фонд документов (информационный	логическое представление информации о хранимой коллекции документов, как о едином хранилище. Содержит индексную и

фонд)	атрибутивную информацию о коллекции документов.
Источник документов (источник данных)	коллекция документов, управляемая единым программным сервером, единой структуры, располагающаяся в сетевой файловой системе или в реляционных базах данных.
Карточка документа	связанная с документом совокупность служебных полей фонда, содержащих информацию о свойствах документа, доступная для просмотра пользователю.
Кластерный анализ списка документов	динамическое группирование списка документов в тематически однородные группы. Используется для выявления макроструктуры списка и погружения в документы кластеров.
Контроллер	отдельная программа, предназначенная для запуска и остановки автоматов на данном компьютере. Транслирует инструкции от менеджера к автоматам.
Концептуальный критерий	абстрактный поисковый запрос для выделения фактов выбранного атрибута, созданный в форме логического высказывания над термами (элементами критерия) без привязки к синтаксису языка запросов источника фактов. Используется для последующего автоматического отображения в физический критерий в диалекте каждого источника.
Критерий	условие на поиск фактов на выбранном языке для фильтра атрибута. Критерий может быть концептуальным или физическим.
Лингвистический фильтр	коллекция отдельных критериев поиска фактов (критериев) на различных языках. Каждый фильтр атрибута соответствует одному хранилищу документов.
Масштабируемость	способность изделия сохранять целостность, устойчивость и производительность при развитии информационных фондов как количественно, так и по составу источников.
Менеджер	отдельная программа, предназначенная для управления диспетчеризации работы автоматов по обработке контроллерами очередей.
Модуль	уровень декомпозиции системы, программа или функционально связанная группа программ.
Объект высказывания	сущность, о которой говорит субъект высказывания в форме прямой или косвенной речи. Например, «Наш начальник сказал: «Петров замечательный специалист» (объект высказывания – «Петров»), «Министр МЧС говорил о том, что проблема выживания России в условиях кризиса стоит наиболее остро» (объект высказывания – «проблема выживания России»).
Объект мониторинга	любая сущность, по которой ведется мониторинг в задачах аналитической обработки (например, физическое или юридическое лицо).

Объект тональности	какой-либо объект, о котором идёт речь в тексте, относительно которого выражена тональность. Например, в предложении За последние две недели экстремисты в Кабардино-Балкарии совершили сразу два громких нападения объектом тональности можно считать экстремистов, а в предложении Госсекретарь США Хиллари Клинтон призвала власти Египта уважать законные права граждан — Хиллари Клинтон.
Очередь	множество идентификаторов документов, отобранных для обработки в рамках задачи.
ПО	программное обеспечение.
Подсистема	программа, рассматриваемая как единое целое, выполняющая законченную функцию и применяемая самостоятельно или в составе комплекса
Поисковая выдача (доставленные документы)	коллекция документов, доставленная клиенту в качестве результата обработки запроса. Ее размер может быть меньше числа найденных по запросу документов.
Пользователь	лицо или организация, использующие действующую систему для выполнения своих задач.
Приоритет задачи	настраиваемый пользователем параметр задачи. Очереди каждой задачи обрабатываются в порядке убывания установленного приоритета.
Программное изделие	программное средство, изготовленное, прошедшее испытания установленного вида и поставляемое как продукция производственно-технического назначения для применения в автоматизированных системах.
Регламент задачи	совокупность временных интервалов, с указанием начала, продолжительности и периодичности, определяющих моменты начала обработки задачи.
Реферат списка документов. Тематический реферат	документ, отражающий основное содержание списка документов. Тематический реферат отражает основное содержание списка документов по выбранной теме.
Рубрика	тематический признак документов, посвященных определенной предметной области.
Синтагма	совокупность нескольких слов, объединённых по принципу семантико-грамматической сочетаемости, единица синтагматики. Объём конкретной синтагмы определяется не только реальным употреблением слов в связке, но и самой возможностью объединения предметов, признаков и процессов окружающей действительности. Минимальной длиной синтагмы следует считать простые словосочетания.

Словарь стоп-тем	общий словарь стоп-тем, не сохраняемых системой. Используется для проверки обрабатываемых тем с целью их не сохранения в справочнике тем
Справочник тем	общий справочник приложения сгруппированных тем-синонимов, в котором расположены активные темы, выделенные и сохраненные для последующего использования. Каждая группа имеет тему – доминанту своего синсета. При сохранении тем производится проверка тем по справочнику и сохраняется ссылка на тему – ее доминанту, если она имеется. Шумовые темы содержатся в словаре стоп-тем (см. *словарь стоп-тем*).
Субъект высказывания	автор высказывания о чем- или о ком-либо, приводимого в тексте в форме прямой или косвенной речи.
Субъект тональности	какой-либо объект – инициатор выраженной в тексте эмоциональной оценки (в общем случае это автор текста). Однако если автор текста ссылается на чьё-нибудь мнение, как в предложении (1) ниже, или цитирует высказывание другого человека, как в предложении (2), то субъектом тональности будет тот, на чьё мнение ссылаются: (1) *Религиоведение, по мнению С. А. Бурьянова, сегодня не представляет собой точной науки, характеризующейся единством и располагающей строгими и общепринятыми принципами* (2) *Глава ЦИК Вешняков вчера в очередной раз похвалил изменения в закон о выборах и сказал, что теперь законодательство перекрывает многие лазейки для злоупотреблений*
Сущность	лингвистически связанная цепочка терминов, выделенная из текста факта и выполняющая функцию его аргумента. Каждая сущность представляет собой типизированный элемент системы.
Тезаурус	особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы и т.п.) между лексическими единицами.
Терминологический вектор документа	упорядоченный по убыванию частоты, веса список тем документа. В качестве веса обычно используется отношение TF (Term Frequency in document) к DF (Term Frequency in all documents).
Тип сущности	в системе выделяются следующие типы сущностей: место и временной отрезок. Для идентификации типа сущности используются специализированные лингвистические

программы.

Тональность — выраженная в тексте эмоциональная оценка. Например, предложение За последние две недели экстремисты в Кабардино-Балкарии совершили сразу два громких нападения. содержит отрицательную оценку происходящего, а предложение Госсекретарь США Хиллари Клинтон призвала власти Египта уважать законные права граждан. – положительную оценку. Тональность высказывания определяется тремя компонентами: субъектом тональности (кто высказал оценку), объектом тональности (о ком или о чём высказана оценка) и собственно тональной оценкой (как оценили).

Факт — типизированное (классифицированное атрибутом) событие (как правило, зафиксированное и произошедшее), дополненное выделенными сущностями, например, временной и пространственной (ссылка на географическое место) метками. К специальному типу сущностей относятся также объекты, упоминаемые в факте. Факт может определять как свойства объекта, так и его связи с другими объектами. Понятие факта характеризуется рядом свойств:

- - тип факта;
- - объект-инициатор факта (субъект);
- - место действия факта;
- - объект-участник факта;
- - время длительности факта;
- - значение факта (например, "продажа бизнеса").
- В модели системы факт F представляется вектором
- F = <ID, A, E, S, text>, где:
- ID – уникальный идентификатор факта,
- A – атрибут, к которому относится факт,
- E – сущности,
- S – множество документов, либо записей, подтверждающих факт,

text – неструктурированное представление (значение) факта.

Физический критерий — критерий, сформированный квалифицированным в области систем управления данными специалистом на языке запросов выбранного источника фактов.

Фильтр атрибута — см. Лингвистический фильтр.

Фонд документов — именованная в системе коллекция документов. Используется в операциях поиска информации. Создается программой User Manager. Может иметь несколько индексов, которые связываются между собой в фонд. Варианты написания названий

элементов списка следующие:

- 1 – список индексов,
- 2 – названия всех индексов фонда представлены шаблоном вида <строка символов> + «*»,
- 3 – комбинация вариантов 1 и 2.

Хранилище документов	программный интерфейс для хранения и поиска документов, а также сами документы из источника документов (см. Источник документов), доступное для поиска и аналитической обработки. Включает в себя средства доступа к полнотекстовому индексу документов.

1 Общие сведения о подсистеме

1.1 Назначение системы

Подсистема управления процессами аналитической обработки предназначена для автоматизированного управления задачами аналитической обработки текста и выделения фактографической информации из структурированных и неструктурированных источников данных.

1.2 Функции подсистемы

Подсистема обеспечивает выполнение следующих функций системы "X-Files":

- установка регламента выполнения задач,
- мониторинг процессов выполнения задач,
- мониторинг состояния компонентов подсистемы (менеджера автоматов, контроллеров и автоматов),
- управление приоритетами задач в режиме реального времени;
- выполнение задач аналитической обработки данных:
  - выделение из документов (информационных сообщений) фактографической информации об изучаемых объектах;
  - накопление выделенной фактографической информации в виде электронных досье на объекты мониторинга;
  - автоматическая типизация объектов и сущностей.

1.3 Требования к аппаратному и программному обеспечению

1.3.1 Требования к аппаратному обеспечению

Минимальные требования к аппаратному обеспечению, которые должны быть выполнены для производительного и устойчивого функционирования подсистемы, приведены в Табл. 1.

Табл. 1 Минимальные требования к аппаратному обеспечению

Название	Минимальные требования
Сервер приложений	– процессорных ядер x64 не менее 2; – оперативная память объемом 12 Гбайт; – жесткий диск объемом 500 Гбайт.
Рабочая станция пользователя	– процессор не ниже Pentium 4; – оперативная память объемом 1 Гбайт.

Примечание – Для снижения времени, затрачиваемого на обработку документов, рекомендуется масштабирование подсистемы. При работе на каждом сервере оптимального числа автоматов (равного числу процессорных ядер сервера), время на

обработку коллекции документов сокращается пропорционально числу серверов. Масштабирование подсистемы за счет увеличения числа автоматов и количества серверов позволяет распараллеливать процессы обработки длинных очередей сообщений.

1.3.2 Требования к общесистемному программному обеспечению

Для функционирования подсистемы должно быть установлено общесистемное программное обеспечение (ОПО), требования к которому приведены в Табл. 2.

Табл. 2 Требования к ОПО

Аппаратное обеспечение	Требования к установленному ОПО
Рабочая станция пользователя:	– операционная система Windows XP Professional с установленным пакетом обновлений Service Pack 3 или выше.
Сервер приложений	– операционная система Windows 2008 Server R2; – СУБД MS SQL Server 2008 R2 или выше; – платформа .NET Framework 4.0; – PROMT Translation Server 8.x Developer Edition (требуется только для обработки многоязычных информационных материалов).

2 Структура подсистемы

Подсистема включает в себя следующие компоненты:

- компонент выполнения процессов аналитической обработки;
- компонент лингвистической обработки текстов на различных языках.

В состав компонента выполнения процессов аналитической обработки входят:

- менеджер автоматов;
- администратор менеджера автоматов;
- контроллеры автоматов, выполняющие запуск автоматов с параметрами задач;
- программы-автоматы (исполняемые модули), которые выполняют задачи аналитической обработки текстов документов для выполнения задач системы "X-Files":
- автомат выделения многоязычной фактографической информации из неструктурированных данных (XFSSQAutomat);
- автомат выделения фактографической информации из структурированных записей БД (DBLoader);

В состав компонента лингвистической обработки текстов на различных языках входят:

- модуль многопоточной лингвистической обработки документов;
- модуль лингвистического анализа текста.

Подсистема обеспечивает единый бизнес-процесс аналитической обработки документов в рамках решения задач системы.

Схема взаимодействия компонентов подсистемы представлена на Рис. 1.

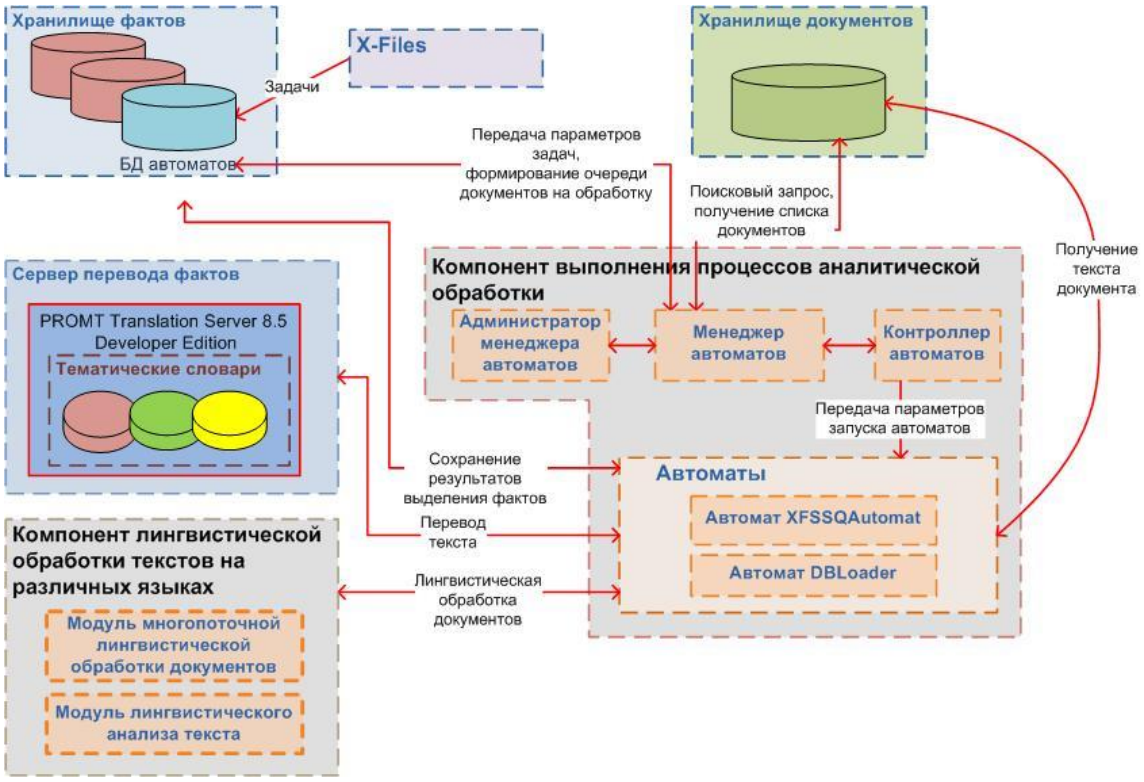

Рис. 1 Схема взаимодействия компонентов подсистемы (в составе системы "X-Files")

2.1 Компонент выполнения процессов аналитической обработки

2.1.1 Администратор менеджера автоматов

Компонент администратора менеджера автоматов является визуальным приложением и выполняет следующие функции:

- управление регламентом выполнения задач;
- отображение состояния контроллеров и автоматов, управляемых менеджером автоматов;
- отображение состояния обработки задач и ошибок, возникших в ходе обработки.

2.1.2 Менеджер автоматов

Компонент менеджера автоматов является управляющей частью компонента и выполняет следующие функции:

- мониторинг статуса задачи (активная/неактивная);
- управление состоянием подключенных контроллеров;
- передача командных инструкций контроллерам;
- запуск или останов задач по регламенту (выполнение регламента);
- выполнение поиска документов, удовлетворяющих критериям, заданным в параметрах задачи и формирование очереди документов на обработку;
- анализ состояния очередей и контроллеров и равномерное распределение вычислительной нагрузки по серверам (запуск автоматов для обработки более критичной задачи на наименее загруженных серверах или приостановка выполнения менее критичных задач).

2.1.3 Контроллер автоматов

Контроллер автоматов является модулем, управляющим работой автоматов подсистемы на выделенном сервере, и выполняет следующие функции:

- мониторинг состояния автоматов, запущенных на сервере контроллера;
- распределение нагрузки между автоматами;
- выполнение инструкций менеджера автоматов по запуску новых и остановке запущенных автоматов.

2.1.4 Автоматы

Автоматы представляют собой самостоятельные программные модули, которые работают под управлением соответствующих контроллеров. Автоматы предназначены для извлечения фактографической информации об объектах из текста многоязычных информационных материалов и формирования досье на целевые объекты.

2.1.4.1 Автомат XFSSQAutomat

Автомат выделения фактов из неструктурированных данных XFSSQAutomat реализует следующие функции:

- автоматическая идентификация объектов путем сопоставления лингвистического описания объекта в системе и в источнике документов;
- предварительная фильтрация фактов в документах с использованием морфологии, операторов булевой алгебры ("И", "ИЛИ", "НЕ"), шаблонов "*" и "?";
- автоматическое выделение из текста фактов по объектам мониторинга с использованием лингвистических фильтров атрибутов, а также с учетом анафорических ссылок между предложениями текста;
- выделение объектов, связанных с целевыми объектами, указанными в задаче;
- автоматический контроль на наличие повторов фактов;
- структурирование фактов, выделение свойств, объектов. Интегрирование (смысловое объединение) выделенных фактов в единое пространство взаимосвязанных досье на основе общих свойств различных фактов: по наименованию объектов, по месту, по времени;
- сохранение выделенных фактов в хранилище фактов.

Исполняемым модулем автомата XFSSQAutomat является *XFSSQAutomat.exe.* Ресурсы для обеспечения работы автомата находятся в следующих папках:

- *LuceneSearchEngine* (файлы поискового движка, обеспечивающего подключение к хранилищу документов Lucene);
- *Logs* (файлы журналов автомата, создается при первом запуске автомата).

2.1.4.2 Автомат DBLoader

Автомат DBLoader реализует процесс автоматического выделения фактографической информации из источников структурированных данных (реляционных БД).

Исполняемым модулем автомата DBLoader является файл *XFAutomat.exe*. Ресурсы для обеспечения работы автомата находятся в следующих папках:

- *Logs* (содержит файлы журнала работы автомата, создается при первом запуске автомата).

Менеджер автоматов выполняет SQL-запрос, заданный в параметрах задачи, к реляционному источнику фактов. Данные, полученные в результате запроса, помещаются в очередь на обработку. Для обработки этих данных запускается автомат DBLoader (или несколько автоматов), который берет очередной необработанный объект, извлекает данные для него из БД. В результате факт на основе этих данных сохраняется в хранилище системной и аналитической информации.

2.2 Компонент лингвистической обработки

В состав компонента лингвистической обработки текстов на различных языках входят:

- модуль многопоточной лингвистической обработки документов;
- модуль лингвистического анализа текста.

Модуль многопоточной лингвистической обработки документов. Модуль обеспечивает выполнение следующих функций:

- доступ автоматов к лингвистическим ресурсам;
- одновременная параллельная обработка массива документов.

Модуль состоит из основного сервиса, который запускает клиентские процессы обработки, формирует очереди документов на обработку для каждого процесса и следит за статусом выполнения этих процессов.

Модуль лингвистического анализа текста. Модуль обеспечивает выполнение следующих функций в рамках лингвистического анализа текста документов:

- лексический анализ (разбиение текста на предложения и лексемы);
- морфологический анализ (определение морфологических характеристик слов, таких как: часть речи, род, число, падеж и т.д.);
- предсинтаксический анализ (выделение групп лексем - синтагм и др.);
- синтаксический анализ (построение дерева разбора предложения и определение синтаксических ролей слов в предложении: подлежащее, сказуемое, дополнение, обстоятельство и т.д.);
- постсинтаксический анализ (выделение объектов и сущностей для типизации);

Последующий семантический анализ текста производит типизацию сущностей (физические, юридические лица; одушевленные предметы; даты; регионы и многие другие типы), а также их нормализацию.

Средства лингвистического анализа текста строго ориентированы на конкретный язык и обслуживаются соответствующим лингвистическим обеспечением, индивидуально разработанным для каждого языкового признака.

При лингвистической обработке текстов на иностранных языках в рамках выполнения задач системы "X-Files" используется программное средство машинного перевода текста Promt Translation Server 8.x Developer Edition.

Лингвистические ресурсы компонента находятся в следующих папках:

- *CommonServices* (содержит общие библиотеки, используемые в лингвистике);
- *Interopes, Libs* (содержат вспомогательные библиотеки)
- *Database* (содержит файлы лингвистических ресурсов для работы автомата – словари аббревиатур, регулярных выражений, географических названий, тональной лексики, морфологические и синтаксические базы);
- *Core* (содержит общие и специализированные лингвистические библиотеки);
- *LingvisticService* (содержит файлы сервиса лингвистики LingvisticService.exe, LingvisticService.exe.config, LingvisticController.dll);
- *UserResources* (содержит словари и другие ресурсы, настраиваемые пользователем).

В файле *ss.ini*, расположенном в папке лингвистического обеспечения, указываются относительные пути к подкаталогам лингвистических ресурсов.

2.3 Взаимодействие компонентов

Программные модули компонента выполнения процессов аналитической обработки развертываются на сервере приложений. Масштабируемость процессов аналитической обработки обеспечивается за счет параллельной работы автоматов на одном, либо на нескольких динамически подключаемых компьютерах. Поэтому при большом количестве задач аналитической обработки рекомендуется использовать несколько серверов приложений.

Сервис менеджера автоматов (см. 2.1.2) взаимодействует с сервисами контроллеров автоматов (см. 2.1.3) на серверах приложений, формирует очереди для каждой активной задачи (см. Определения и сокуращения) и распределяет вычислительную нагрузку по очередям.

Каждый контроллер автоматов получает от менеджера задач идентификаторы документов очереди, запускает автоматы. Автоматы обрабатывают документы очереди в соответствии со своими функциями (см. 2.1.4).

Выполнение задач производится в соответствии с расписанием задач и их приоритетом. Выполнение каждой задачи может быть приостановлено, затем продолжено с того места, где обработка была закончена.

В процессе выполнения задачи автоматы обращаются к компоненту лингвистической обработки текстов на различных языках, выполняющему морфологический, синтаксический и семантический анализ текста.

Процедура создания и обработки задач включает следующую последовательность действий:

1. В интерфейсе системы создается задача аналитической обработки.
2. В приложении администратора менеджера автоматов создается задача менеджера автоматов и настраиваются параметры выполнения задачи:
   - временной регламент запуска задачи;
   - приоритет выполнения;
   - период пополнения очереди.
3. Менеджер автоматов:
   - считывает параметры задачи на выполнение;
   - выполняет поисковый запрос к хранилищу документов, удовлетворяющий условиям задачи, формирует список документов и сохраняет его в очереди;
   - направляет контроллерам автоматов команды управления задачами (останова, запуска, сброса);
   - периодически опрашивает узлы контроллеров автоматов.
4. Контроллер автоматов в соответствии с приоритетом задачи запрашивает список документов очереди и запускает автоматы для обработки очереди.
5. Автоматы обрабатывают документы очереди. Результаты обработки сохраняются в хранилище системной и аналитической информации.

Все результаты аналитической обработки просматриваются через веб-приложения системы.

3 Установка и настройка подсистемы

3.1 Процедура автоматической установки

Для установки компонента необходимо запустить командный файл *QM.exe* из каталога "..\Инсталляционный комплект\Install". На экране отображается стартовая страница установки (Рис. 2).

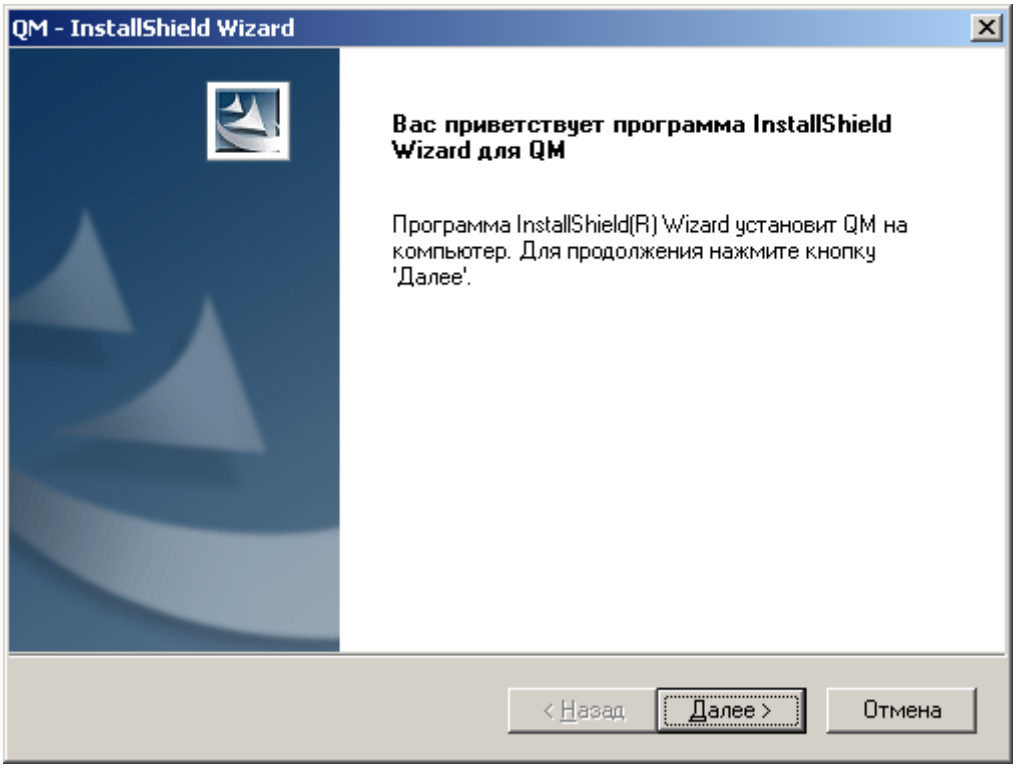

Рис. 2 Стартовая страница установки компонента

Необходимо нажать кнопку "Далее". На экране отобразится окно для ввода сведений о пользователе (Рис. 3).

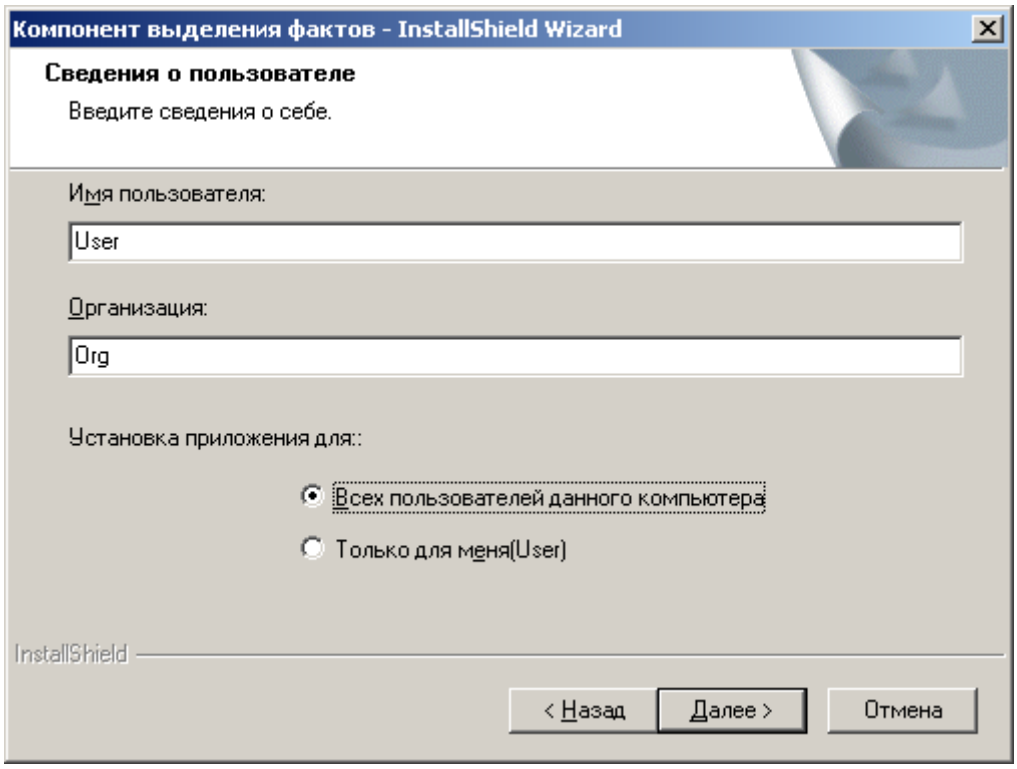

Рис. 3 Окно для ввода сведений о пользователе

В этом окне необходимо заполнить все предлагаемые поля, оставить поле "Установка приложения для всех пользователей данного компьютера" и нажать кнопку "Далее".

Если необходимо вернуться к предыдущему этапу и изменить настройки, то следует нажать кнопку "Назад" (то же для всех последующих окон).

На экране отобразится окно для выбора вида установки (Рис. 4).

При выборе вида установки "Обычная" (по умолчанию) и нажатии на кнопку "Далее" на экране отобразится окно, показывающее ход выполнения установки (Рис. 7).

При выборе вида установки "Выборочная" и нажатии на кнопку "Далее" на экране отобразится окно для выбора папки, в которую будут установлены файлы (см. далее Рис. 5).

Вид установки "Сокращенная" предназначен для установки компонентов на рабочую станцию администратора.

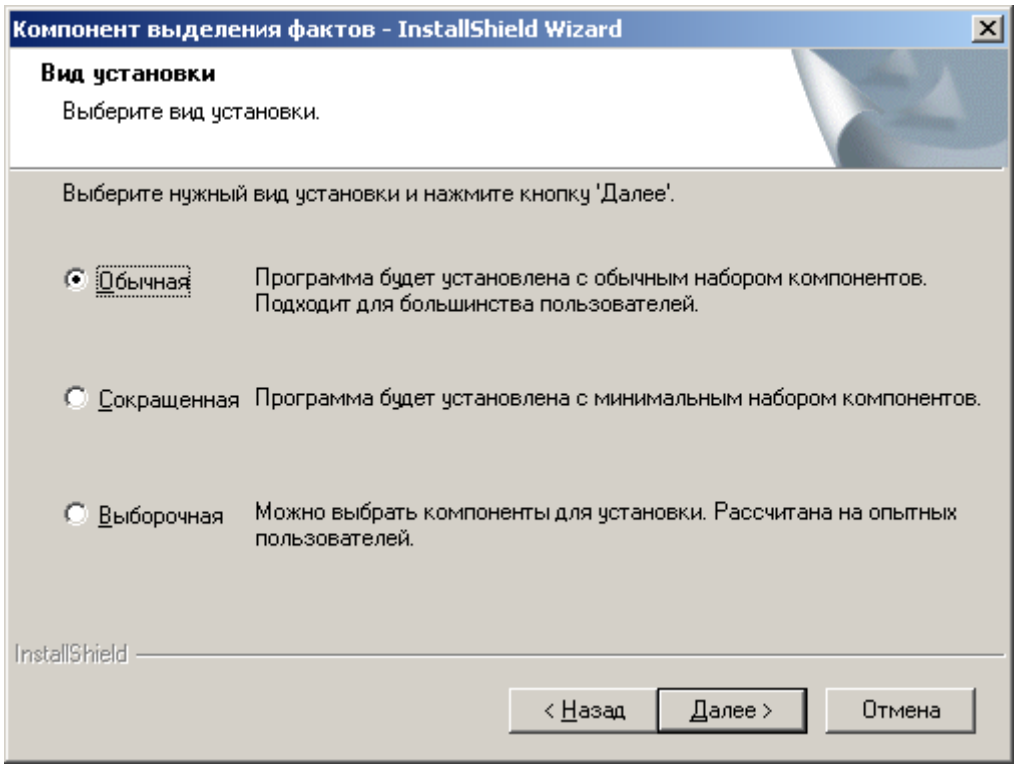

Рис. 4 Окно для выбора вида установки

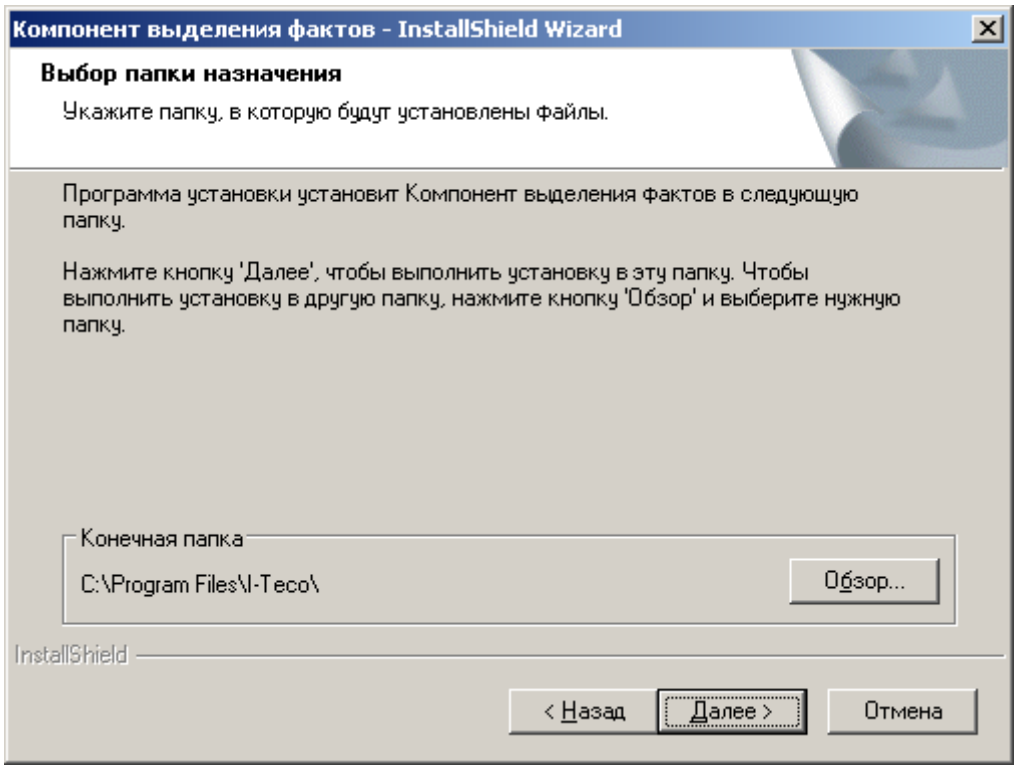

Рис. 5 Окно для выбора папки, в которую будут установлены файлы

После выбора папки и нажатия кнопки "Далее" отобразится окно выбора необходимых компонентов для установки (Рис. 6). Данный вид установки позволяет выбрать компоненты в зависимости от конфигурации сервера.

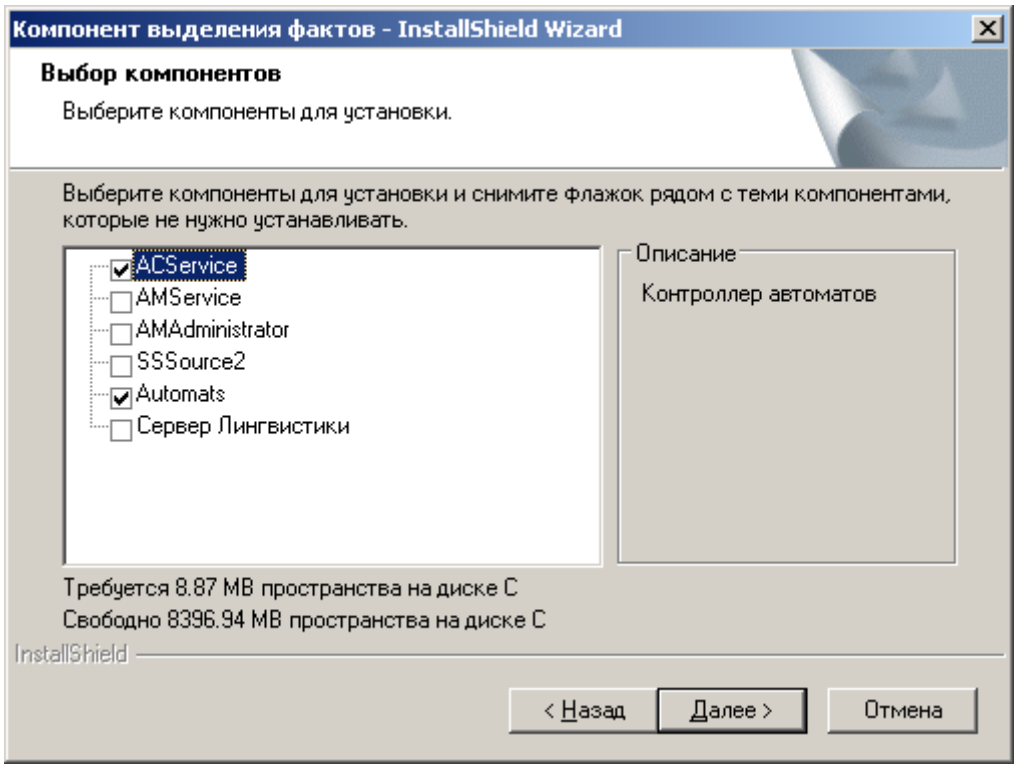

Рис. 6 Окно выбора компонентов для установки

В этом окне следует выбрать компоненты для установки и нажать кнопку "Далее". На экране отобразится окно, показывающее состояние установки (Рис. 7).

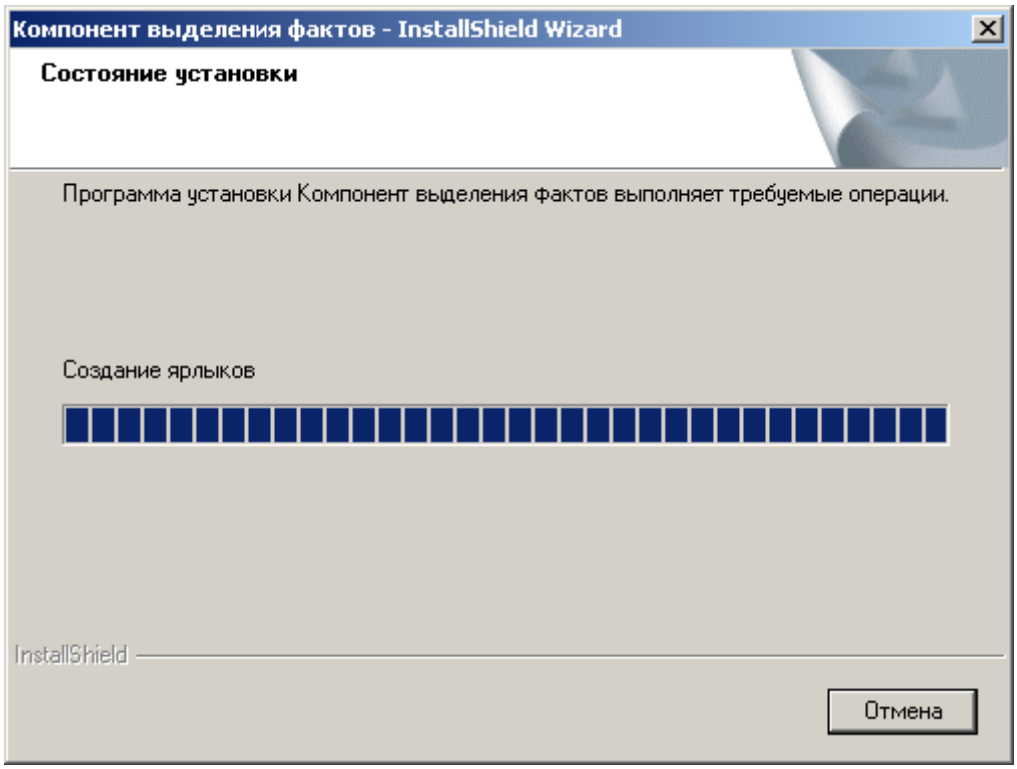

Рис. 7 Окно состояния установки

После окончания копирования на экране отобразится сообщение об успешной установке, представленное на Рис. 8.

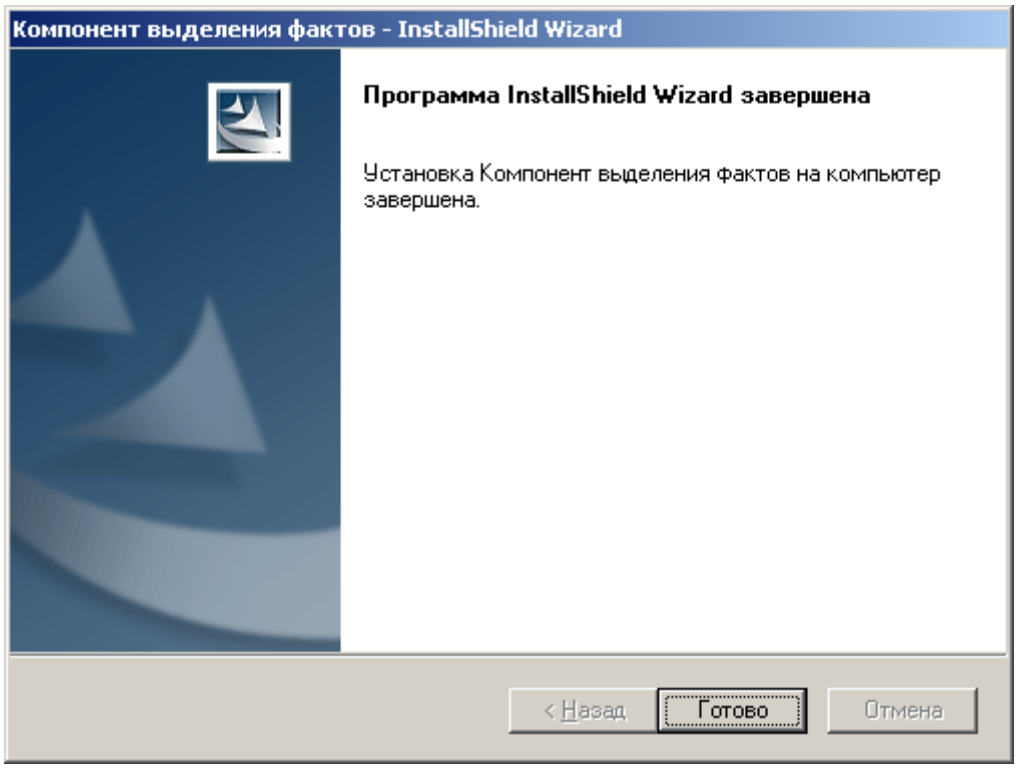

Рис. 8 Сообщение об успешной установке компонента

В процессе установки (если выбран вариант установки "Обычная") выполняются следующие действия:

- Устанавливается компонент выполнения процессов аналитической обработки, в том числе:
  - менеджер автоматов AMService;
  - контроллер автоматов ACService;
  - администратор менеджера автоматов AMAdministrator;
  - автомат XFSSQAutomat;
  - автомат XFDBLoader.
- Устанавливается компонент лингвистической обработки текстов на различных языках, в том числе сервис лингвистики LingvisticServer.

После окончания установки необходимо настроить модули компонентов (см. разделы 3.4, 3.5).

3.2 Процедура ручной установки

3.2.1 Установка компонента процессов аналитической обработки

В данном разделе описана процедура установки программных модулей компонента выполнения процессов аналитической обработки.

3.2.1.1 Установка администратора менеджера автоматов

Для установки приложения администратора менеджера автоматов необходимо скопировать папку AMAdministrator на установочном диске системы в целевой каталог установки.

3.2.1.2 Установка менеджера автоматов

Для установки службы менеджера автоматов необходимо скопировать папку AMService на установочном диске системы в каталог установки и отредактировать файл *install.bat.* В файле следует указать корректный путь к утилите InstallUtil.exe для MS .Net Framework 4.0 в строке по шаблону:

>[путь к утилите для framework 4.0]\installUtil.exe AMService.exe.

По умолчанию, утилита InstallUtil.exe расположена в каталоге *C:\WINDOWS\Microsoft.NET\Framework\v4.0.30319*.

При успешной установке выдается сообщение: *"The Commit phase completed successfully"*, что завершает установку сервиса менеджера.

Далее необходимо выполнить настройку сервиса AMService (см. раздел 3.4.2).

3.2.1.3 Установка контроллера автоматов

Для установки сервиса контроллера автоматов необходимо скопировать папку ACService на установочном диске системы в каталог установки и отредактировать файл *install.bat.* В файле следует указать корректный путь к утилите InstallUtil.exe для MS .Net Framework 4.0 в строке по шаблону:

>[путь к утилите для framework 4.0]\installUtil.exe ACService.exe.

По умолчанию, утилита InstallUtil.exe расположена в каталоге *C:\WINDOWS\Microsoft.NET\Framework\v4.0.30319*.

При успешной установке выдается сообщение: *"The Commit phase completed successfully"*, что завершает установку сервиса контроллера.

Далее необходимо выполнить настройку сервиса ACService (см. раздел 3.4.3).

3.2.1.4 Установка и настройка автоматов

Для установки автоматов аналитической обработки достаточно с копировать каталоги автоматов на установочном диске в целевой каталог установки.

Далее для каждого автомата необходимо выполнить настройку параметров работы (см. раздел 3.4.5).

3.2.2 Установка компонента лингвистической обработки текстов на различных языках

При установке компонента файлы лингвистического обеспечения, необходимые для работы автоматов, устанавливаются в папку Lingvistic целевого каталога установки.

Для установки сервиса лингвистической обработки необходимо в папке .\Lingvistic\LingvisticService\ отредактировать файл *install.bat* и запустить его на выполнение. В файле следует указать корректный путь к утилите InstallUtil.exe для MS .Net Framework 4.0 в строке по шаблону:

>[путь к утилите для framework 4.0]\installUtil.exe ACService.exe.

По умолчанию, утилита InstallUtil.exe расположена в каталоге *C:\WINDOWS\Microsoft.NET\Framework\v4.0.30319*.

При успешной установке выдается сообщение: *"The Commit phase completed successfully"*, что завершает установку сервиса лингвистической обработки.

Далее необходимо выполнить настройку сервиса LingvisticService (см. раздел 3.5.3).

3.3 Настройка общесистемного ПО

Для настройки взаимодействия сервиса PROMT с автоматами подсистемы необходимо создать пользователей и группы пользователей веб-сервера PROMT. Для этого требуется:

– Загрузить веб-страницу http://localhost/pts8 на сервере, где установлен PROMT;

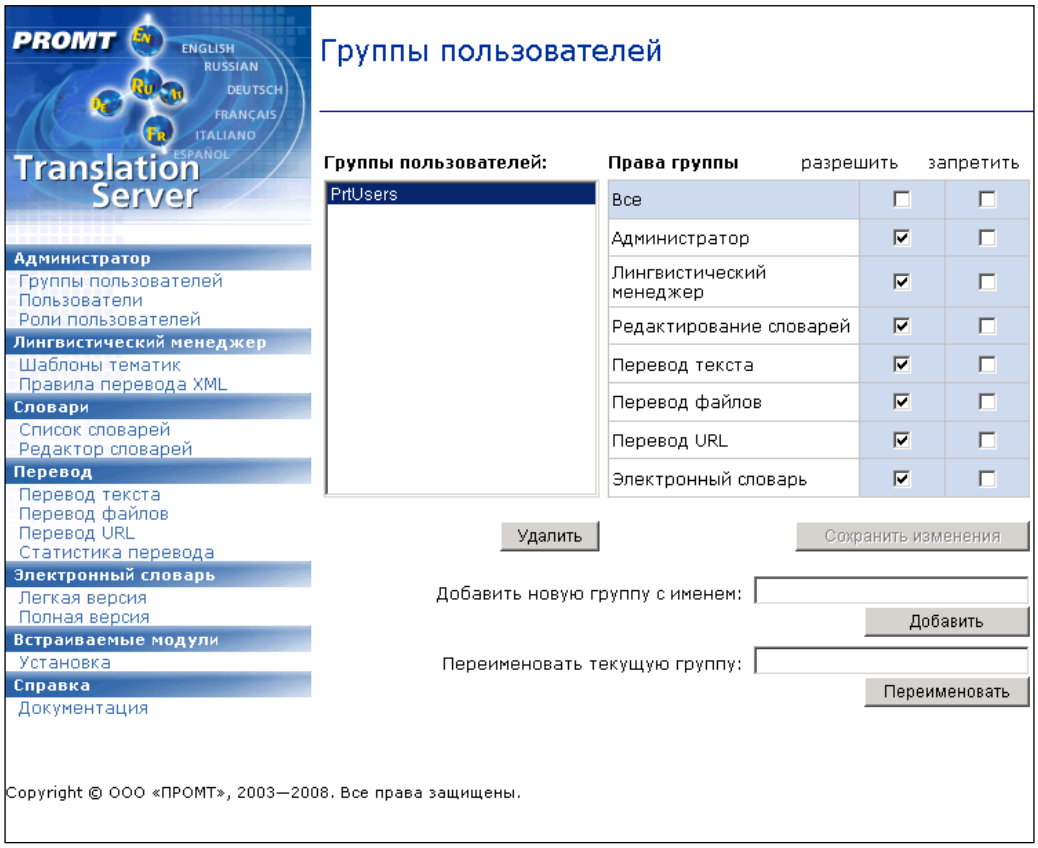

Рис. 9 Настройки рабочих групп PROMT

– Зайти в раздел "Администратор"-"Группы пользователей" и создать группу PrtUsers с административными правами (см. Рис. 9);
– Создать пользователя Windows "promt" и установить пароль Windows для него с неограниченным сроком;

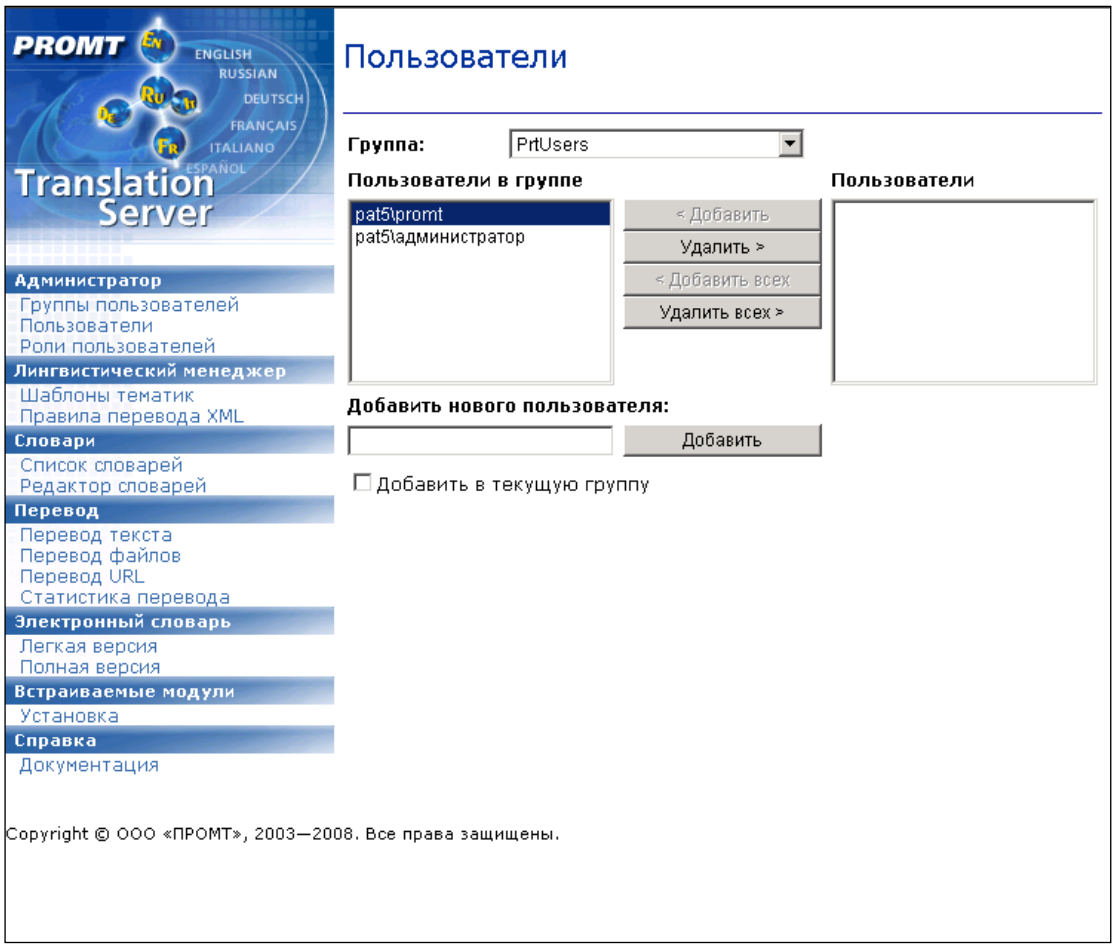

Рис. 10 Настройки пользователей PROMT

– Зайти в раздел "Администратор"-"Пользователи" и добавить новых пользователей с административными правами в созданную группу (Рис. 10). Необходимо указать имя пользователя в формате <сервер>\<имя>.

3.4 Настройка компонента процессов аналитической обработки

3.4.1 Настройка администратора менеджера автоматов

Для настройки программы администратора менеджера автоматов используется конфигурационный файл AMAdministrator.exe.config. Пример содержимого файла представлен ниже:

```
<userSettings>
    <AMAdministrator.Properties.Settings>
      <setting name="AutoUpdateInterval" serializeAs="String">
        <value>5</value>
      </setting>
      <setting                name="DisplayTimeSplashScreenITeco"
serializeAs="String">
        <value>0</value>
```

```
            </setting>
            <setting                          name="MaxCountDisplayedErrors"
  serializeAs="String">
              <value>1000</value>
            </setting>
            <setting name="ServerName" serializeAs="String">
              <value><server>:<port></value>
            </setting>
          </AMAdministrator.Properties.Settings>
        </userSettings>
```

Параметр `AutoUpdateInterval` задает период автоматического обновления данных (в секундах) в окнах приложения AMAdministrator.

Параметр `DisplayTimeSplashScreenITeco` – время показа стартовой заставки приложения. Время указывается в секундах. Значение "0" – заставка не показывается.

Параметр `MaxCountDisplayedErrors` – максимальное количество отображаемых ошибок в интерфейсе приложения AMAdministrator.

Параметр `ServerName` задает адрес и порт подключения к windows-сервису AMService. Параметр <server> соответствует имени или IP-адресу компьютера, на котором установлен AMService. Параметр <port> соответствует номеру порта, который прослушивает сервис (оба параметра настраиваются в 3.4.2).

3.4.2 Настройка менеджера автоматов

Для настройки необходимо указать в конфигурационном файле AMService.exe.config в установочной папке менеджера автоматов корректные значения параметров. Описание параметров см. в Табл. 3.

Табл. 3 Параметры XML-файла конфигурации менеджера автоматов

Ключ	Назначение	Значение по умолчанию
MainConnectString	Строка подключения к БД менеджера автоматов. Возможно использование поставщика данных OLEDB либо Oracle, в последнем случае строка соединения дополняется ключом "DbClient=Oracle;".	Provider=SQLOLEDB.1; Initial Catalog= AMDB;Data Source=<server>;User ID=<user>; Password=<passwd>;
TaskControlInterval Min	Период (в минутах) сканирования расписания задач и отработка соответствующих действий.	20
UnlockQueueItemTimeout Min	Период (в минутах) запуска процесса снятия блокировок у просроченных документов.	Указано в параметре контроллера автоматов Service:ObjectUri, см. Табл. 4

Ключ	Назначение	Значение по умолчанию
AutomatControllerUri	Строка названия сервиса контроллера автоматов для сетевого доступа.	AutomatController
Раздел <AKAccessor.Properties.Settings>		
EngineName	Название типа поискового движка	"LuceneSearchEngine" для Lucene
EngineParams	Строка параметров для поискового движка (специфична для каждого типа движка)	"IndexPath=<путь к каталогу индекса Lucene>; ConnectorPort=2051; ConnectorUri=LuceneConnector;MaxClauseCount=1000; LingvisticServer=<server>:45000" для Lucene. В параметре LingvisticServer указывается адрес сервера лингвистики
Раздел <system.runtime.remoting>:<application>		
Service:objectUri	Строка названия сервиса менеджера автоматов для сетевого доступа	AutomatManager
Channel:port	Номер порта для доступа к сервису менеджера автоматов из модуля администратора.	44444

Примечание: Изменения вступают в силу после перезапуска сервиса менеджера AMService.

В каталоге менеджера автоматов (по умолчанию, .\AMService) при первом запуске сервиса создается папка *Logs*. В папку *Logs* записываются файлы журналов работы сервиса.

Дополнительно, для каждой задачи выделения фактов и задачи определения тональности создается отдельный файл журнала По умолчанию, для задач выделения фактов создаются файлы XFSSQ_<task id>.log, а для задач определения тональности – файлы AKTon_<task id>.log (где task_id – идентификатор задачи в системе). Журнал содержит данные по выполнению поисковых запросов задачи менеджером автоматов.

3.4.3 Настройка контроллера автоматов

Для настройки необходимо указать в конфигурационном файле ACService.exe.config в установочной папке менеджера автоматов корректные значения параметров. Описание параметров см. в Табл. 4.

Табл. 4 Параметры XML-файла конфигурации контроллера автоматов

Ключ	Назначение	Значение по умолчанию
AMServer	Имя сервера, на котором работает менеджер автоматов	Указано в параметре MainConnectString менеджера автоматов, см. Табл. 3
AMPort	Номер порта для соединения с менеджером автоматов	44444
AMUri	Строка названия сервиса менеджера автоматов для сетевого доступа	Указывается в параметре objectUri менеджера автоматов, см. Табл. 3
ACPort	Номер порта для получения сообщений от менеджера, должен совпадать со значением параметра <system.runtime.remoting><application> <channels> <channel>: port	40040
MaxAutomatCount	Количество одновременно работающих автоматов на данном узле	1
Раздел <AutomatTypeList>:<ArrayOfAutomatType>:<AutomatType>: содержит описание списка автоматов, которые могут быть запущены на данной машине. Каждая секция описания автомата ограничена тегом <AutomatTypeItem>		
ID	Идентификатор типа автомата	Одно из значений: AKSemnet, AKTon, AKTonLoader, XFSSQ, XFDBLoader
FileName	Имя файла модуля автомата	<путь к каталогу автомата>\<имя исполняемого модуля автомата>
WorkingDirectory	Рабочий каталог для модуля автомата	<путь к каталогу автомата>
Раздел <system.runtime.remoting>:<application>		
Service:ObjectUri	Строка названия сервиса контроллера автоматов для сетевого доступа	AutomatController
Channel:port	Номер порта для доступа к сервису контроллера автоматов из менеджера автоматов	Должен совпадать со значением параметра ACPort

Примечание: Изменения вступают в силу после перезапуска сервиса контроллера ACService.

В каталоге контролллера автоматов (по умолчанию, .\ACService) при первом запуске сервиса создается папка *Logs*. В папку записываются файлы журналов работы сервиса.

3.4.4 Настройка автомата XFSSQAutomat

Для настройки автомата используется конфигурационный файл XFAssqEx.exe.config в папке автомата. Описание параметров настройки приведено в Табл. 5.

Табл. 5 Параметры XML-файла конфигурации автомата XFSSQAutomat

Ключ	Назначение	Значение по умолчанию
Раздел <AutomatModule.Properties.Settings>		
QueueScanPeriod	Период сканирования пустой очереди в секундах	60
ProcTimeOut	Максимальное время в секундах для непрерывной обработки одного элемента очереди, в случае превышения данного периода процесс обработки элемента очереди прерывается.	600
Раздел <AutomatModule.XFSSQAutomat.Properties.Settings>: содержит специфические параметры автомата		
EngineName	Название поискового движка, используемого автоматом (см. Табл. 3)	См. Табл. 3, одноименный параметр
EngineParam	Строка параметров для поискового движка (см. Табл. 3)	См. Табл. 3, одноименный параметр
XFLogin	Логин пользователя системы XFiles	XFA
MinTextSizeForSimpleProc	Минимальный размер текста документа для использования упрощенной процедуры лингвистической обработки (исключительно обработка предложений с поисковыми хитами).	10000
PTSUrl	Адрес сервиса перевода	http://<server>/pts8 Указывается при настройке Промт (см. 3.3)
PTSUserName	Имя пользователя для доступа к сервису перевода	Указывается при настройке Промт
PTSPassword	Пароль для доступа к сервису перевода	Указывается при настройке Промт
PTSEnable	Флаг включения функции перевода	True, если установлен Promt False, если Promt отсутствует
LingvisticServer	Адрес и порт подключения к сервису лингвистики	Указывается в параметре Channel:port сервиса лингвистики, см. Табл. 6
UseGeoNameDictionary	Параметр использования справочника для проверки географических названий сущностей, выделяемых из фактов (см. 3.5.1)	True (проверка включена) False для отмены проверки названий

Дополнительная настройка выделения объектов определенного типа из фактов проводится с помощью словаря SemanticSuffixes.dct (находится в каталоге *Database\CoSeDiRussian\Semantics*). Данный словарь содержит список слов, по наличию

которых объект выделяется как юридическое лицо. Словарь содержит такие слова как "связь", "транс", "имущество", "суд", "банк" и позволяет выделять такие юридические лица, как "Москомимущество", "Мосгорсуд", "Внешэкономбанк" и т.д. Словарь имеет простую текстовую структуру и может быть расширен.

3.4.5 Настройка автомата XFDBLoader

Для настройки автомата используется конфигурационный файл XFAutomat.exe.config в папке автомата. Параметры аналогичны перечисленным в Табл. 5.

3.5 Настройка компонента лингвистической обработки текстов на различных языках

При установке компонента файлы лингвистического обеспечения, необходимые для работы автоматов, устанавливаются в папку Lingvistic целевого каталога установки.

Необходимо настроить общий сетевой доступ к каталогу лингвистических ресурсов Lingvistic. Для этого следует в свойствах папки на вкладке "Безопасность" добавить пользователя NETWORK SERVICE и указать ему уровень разрешений "Полный доступ".

3.5.1 Настройка справочника географических названий

Справочник географических наименований используется для проверки принадлежности имени собственного к географическим наименованиям: если имя собственное совпадает с географическим наименованием из справочника, то оно будет помечено как географическое наименование.

Справочник географических наименований представляет собой текстовый файл с именем GeoNames.data и хранится в папке *Lingvistic* в каталоге *.\UserResources\GeographicNames*. Текстовый файл справочника имеет кодировку Unicode. Текст разбит на строки, так что одно географическое наименование занимает одну строку. При редактировании строк и добавлении новых допускается использовать только заглавные буквы и цифры, удаляя пробелы и прочие символы. Например, город Москва в справочнике должен выглядеть как МОСКВА, а город Нью-Йорк — как НЬЮЙОРК по-русски и NEWYORK по-английски.

Для сравнения названий с анализируемыми объектами необходимо создать индекс справочника — файл с именем GeoNames.index, который хранится в той же папке. Для этого используется утилита IndexingGeoNames.exe в папке *.\utils\IndexingGeoNames* каталога Lingvistic. В качестве параметра необходимо указать полный путь к рабочему каталогу, в котором содержится файл GeoNames.data. Например, если каталог *Lingvistic* хранится на локальном диске D, то командная строка для вызова утилиты индексации будет выглядеть так:

```
> IndexingGeoNames.exe D:\Lingvistic\UserResources\GeographicNames
```

При каждом изменении файла GeoNames.data справочник необходимо переиндексировать указанным способом.

По завершении индексации справочник готов к работе, никаких дополнительных действий для его подключения к системе не требуется.

3.5.2 Настройка тональных словарей

Для встроенных семантических типов системы (физические лица, юридические лица, прочие) поддерживаются словари тональной лексики. Словари для каждого семантического типа расположены в соответствующих подпапках в папке .\UserResources\ToneDicts, расположенной в папке Lingvistic. Словари должны быть доступны для редактирования выделенным пользователям.

В словарях задаются все словосочетания, имеющие тональную окраску (позитивную или негативную). Каждое словосочетание характеризуется числом, задающим силу тональности данного словосочетания. Степень негатива задается числом от -10 до -1, а позитива - от 1 до 10.

Файлы словарей являются простыми текстовыми файлами. Каждое словосочетание в файле находится на отдельной строке. При этом за словосочетанием через символ табуляции следует его тональный вес.

В папке словарей каждого семантического типа содержатся следующие словари (название файла словаря совпадает с типом словаря):

- прилагательные-определения (tone_adjectives) - сохраняются прилагательные и наречия, например, "хороший", "плохой", "хорошо", "плохо";
- существительные (tone_nouns) - сохраняются существительные в словосочетаниях, например, "Иванов – негодяй";
- существительные в корне генетивной цепочки (tone_nounsgenerative) - сохраняются существительные в генетивных цепочках, например, "успех президента", "ошибка губернатора";
- существительные - второстепенные члены предложения (tone_objectnouns) - сохраняются любые прочие существительные, встречающиеся в любых других местах предложений как второстепенные члены;
- стоп-слова (tone_stop_words);
- глаголы (tone_verbs) - сохраняются глаголы и глагольные словосочетания.

Для достижения максимальной точности и полноты рубрицирования документов необходимо редактирование словарей тональной лексики. Можно выполнять со словарями следующие действия:

- добавить словосочетания;
- изменить степени тональности для уже созданного словосочетания;
- удалить словосочетания;
- копировать словосочетания в словари других семантических типов и ролей (поскольку одна и та же лексика часто дублируется в словарях разных семантических типов и ролей);
- перенести все значения словаря данной семантической роли и типа в словарь аналогичной роли, но других типов (т.е. можно наполнить словарь

определений, например, для физических лиц, а потом все словосочетания из него перенести в словарь для юридических лиц).

3.5.3 Настройка сервиса лингвистики

Настройка сервиса лингвистики выполняется в конфигурационном файле сервиса LingvisticService.exe.config. Настраиваемые параметры перечислены в Табл. 6.

Табл. 6 Параметры XML-файла конфигурации сервиса лингвистики

Ключ	Назначение	Значение по умолчанию
ProcessorPath	Полный путь к файлу LingvisticProcess.exe	Определяется при установке компонента, см. п. 3.1
StartPort	Начало интервала портов для процессов	40040
EndPort	Конец интервала портов для процессов	40040
MaxProcessorCount	Максимальное число одновременно работающих процессов лингвистической обработки. Другие вызовы на обработку ставятся в очередь	20
ProcessorInitTimeOutInSeconds	Время на инициализацию клиентского процесса	180
ProcessorCloseTimeOutInSeconds	Таймаут на закрытие клиентского процесса	60
ProcessorQueueTimeOutInSeconds	Максимальное время на обработку элемента очереди клиентским процессом. При его превышении обработка элемента останавливается	180
StartupProcessorCount	Начальное число процессов лингвистической обработки при старте сервиса. Значение не должно превышать значения параметра MaxProcessorCount	2
ProcessorIdleTimeOutInSeconds	Таймаут для определения неработающего процесса. При превышении времени процесс считается незанятым и может быть использован для вызовов	600
ProcessorWorkTimeOutInSeconds	Таймаут для вызовов функций лингвистической обработки текста (в сек.)	180
Trace	Включение логирования обработки	False
Раздел <system.runtime.remoting>:<application>:<channels>:		
Channel: port	Номер порта для доступа к сервису лингвистики из приложений	44444

В каталоге сервиса лингвистики (по умолчанию, .\Lingvistic\LingvisticService) при первом запуске сервиса создается папка *Logs*. В папку записываются файлы журналов работы сервиса. В журнале используются следующие показатели времени:

- GetItem – время ожидания документа в очереди на обработку;
- Fun – работа лингвистики.

4 Проверка работы подсистемы

Для проверки работы подсистемы необходимо выполнить следующие действия:

1. На компьютере администратора запустить сервис менеджера автоматов.
2. Запустить сервис контроллера автоматов на каждом из серверов. При использовании режима старта "Auto" необходимо убедиться, что статус сервиса имеет значение "Работает"
3. На компьютере администратора запустить приложение администратора менеджера автоматов. Проверить статус контроллеров в узлах системы автоматов. Описание работы с интерфейсом приложения см. в документе "ДШСК.50 8120 9.003-04 34 04. Руководство оператора".
4. Создать задачу выделения фактов в веб-интерфейсе X-Files.
5. Создать источник задач в приложении администратора менеджера автоматов, соответствующий п.4. Создать задачу для менеджера автоматов, соответствующую п.4. Определить регламент запуска задачи как однократный. Менеджер автоматов будет автоматически выполнять задачу в соответствии с регламентом.
6. Проверить процесс выполнения задач. Если все задачи, которым необходимо запуститься по расписанию, уже запущены и выполняются, то сервисы и модуль администрирования работоспособны. В противном случае необходимо просмотреть журналы тех задач, которые не выполняются, хотя по расписанию должны работать и выявить ошибки, из-за которых они были остановлены.

Ошибки могут возникать по следующим причинам:

– ошибки работы с БД, указанных в параметрах задач. Возможные причины: неправильно установленные параметры соединения с БД, см. п. 4;

– ошибки работы с хранилищем документов. Возможные причины: задача работала в тот момент, когда поисковый сервер выполнял задачи, требующие блокировки хранилища документов (например, задачи резервного копирования), или информационный фонд не найден.

5 Сообщения системному программисту

Табл. 7 Сообщения журнала автомата

Описание журнала	Описание сообщения и способ решения проблемы
В процессе обработки задачи создается журнал в виде текстового файла. Каждая строка журнала – описание того или иного действия, совершенного автоматом: поиск документов по атрибуту и количество найденных, выделение факта, идентификация уже существующего объекта, добавление нового объекта. Строка начинается с даты и времени совершения действия, описанного в строке. Далее следует собственно описание действия.	В случае возникновения ошибки в журнал пишется строка следующего формата: "Дата/время ERROR. <*Описание ситуации*>. <*Описание ошибки*>". Необходимо передать содержимое раздела разработчику системы.

Табл. 8 Сообщения журнала менеджера автоматов

Описание журнала	Описание сообщения и способ решения проблемы
В процессе работы менеджера автоматов создается журнал в виде текстового файла. Каждая строка журнала содержит операции по управлению контроллерами и автоматами. Для каждой задачи создается журнал процесса обработки задачи, содержащий описание действия по управлению выполнением задач - поиск документов по атрибуту и количество найденных, формирование очереди. Каждая строка журнала начинается с даты и времени совершения действия, описанного в строке. Далее следует описание действия.	В случае возникновения ошибки в журнал пишется строка следующего формата: "Дата/время ERROR. <*Описание ситуации*>. <*Описание ошибки*>". Необходимо передать содержимое раздела разработчику системы.

Табл. 9 Сообщения журнала контроллера автоматов

Описание журнала	Описание сообщения и способ решения проблемы
В процессе работы контроллера автоматов создается журнал в виде текстового файла. Каждая строка журнала содержит операции по выполнению команд запуска и остановки автоматов. Каждая строка журнала начинается с даты и времени совершения действия, описанного в строке. Далее следует описание действия.	В случае возникновения ошибки в журнал пишется строка следующего формата: "Дата/время ERROR. <*Описание ситуации*>. <*Описание ошибки*>". Необходимо передать содержимое раздела разработчику системы.

Приложение 1

Настройка тональных словарей модуля разметки тональности

1. Структура тональных словарей

Модуль тональности содержит 26 словарей тональной лексики. Каждый словарь представляет собой текстовый файл в кодировке UTF-8 с расширением "txt". Словари разбиты на классы по частям речи и добавлены списки коллокаций [1]. Всего определено пять классов (см. Табл. 10):

- **Nouns** - существительные
- **Verbs** - глаголы
- **Adjectives** - прилагательные
- **Advrebs** - наречия
- **Collocations** – коллокации.

Все словари представляют собой списки слов (одно слово в строке), за исключением ***Collocation и Verbs***. В ***Collocation*** возможны словосочетания из двух и более слов. В ***Verbs*** возможны слова и словосочетания из двух слов (глагол + управление). В одном классе частей речи одно и то же слово не может встретиться дважды.

Табл. 10 Классификация тональных словарей

Класс	Имя файла	Обозначение	Идентификатор
Nouns	list_nouns_aneg.txt	action negative	nANeg
	list_nouns_apos.txt	action positive	nAPos
	list_nouns_neg.txt	negative	nNeg
	list_nouns_pos.txt	positive	nPos
	list_nouns_pneg.txt	potential negative	nPNeg
	list_nouns_ppos.txt	potential positive	nPPos
Verbs	list_verbs_flxneg.txt	reflexive negative	vFlxNeg
	list_verbs_flxpos.txt	reflexive positive	vFlxPos
	list_verbs_link.txt	related verbs	vLink
	list_verbs_neg.txt	negative	vNeg
	list_verbs_pos.txt	positive	vPos
	list_verbs_pure_neg.txt	negative pure	vNegP
	list_verbs_pure_pos.txt	positive pure	vPosP
	list_verbs_opp_pos.txt	opposition positive	vOppPos
	list_verbs_opp_neg.txt	opposition negative	vOppNeg
Adjectives	list_ adjectives_neg.txt	negative	adjNeg
	list_adjectives_pos.txt	positive	adjPos

[1] Под коллокациями понимались любые устойчивые и достаточно часто встречающиеся сочетания слов, как с идиоматическим значением, так и с неидиоматическим.

Класс	Имя файла	Обозначение	Идентификатор
	list_adjectives amplf.txt	amplificator	adjAmplf
Adverbs	list_adverbs_neg.txt	negative	advNeg
	list_adverbs_pos.txt	positive	advPos
	list_adverbs amplf.txt	amplificator	advAmplf
Collocations	list_collocation_neg.txt	negative	collNeg
	list_collocation_pos.txt	positive	collPos
	list_collocation_amplf.txt	amplificator	collAmplf
	list_collocation_vneg.txt	negative	collvNeg
	list_collocation_vpos.txt	positive	collvPos

В целом, все словари можно разделить на две большие группы: неглагольные лексемы и коллокации и глагольные лексемы и коллокации.

2. Неглагольные лексемы и коллокации

Наречия (*adv - adverbs*), прилагательные (*adj - adjectives*) и неглагольные коллокации (*coll - collocations*) делятся на позитивные (*Pos - positive*), негативные (*Neg - negative*) и усиливающие эмоциональность (*Amplf - amplificators*), то есть такие слова или словосочетания, которые сами по себе не несут тональности, но при этом могут усиливать эмоциональность того, к чему присоединяются (например, наречия *круто, ужас*; прилагательные *эксклюзивный, потрясающий* и коллокации *коренным образом, решающая роль*).

Имена существительные (*n - nouns*) также могут быть позитивными (например, *благотворительность* или *зарплата*) и негативными (*налог* или *война*). Однако не все существительные имеют однозначную эмоциональную нагрузку, тональность многих зависит от окружения. Поэтому введены классы потенциально негативных (*nPNeg - potential negative*) и потенциально позитивных (*nPPos - potential positive*) слов — так, потенциально позитивные слова позитивны в позитивном окружении и нейтральны во всех остальных. Например, слово *план* само по себе не несёт в себе тональности, но сочетание *план по выходу из кризиса* должно давать позитив.

Особую роль играют отглагольные существительные: они могут менять тональность следующего за ним существительного. Например, отглагольное существительное *прекращение* меняет её на противоположную. Если за ним следует позитивная цепочка связанных существительных, например, *прекращение поставок угля*, то объединенная цепочка будет негативной. Если за данным отглагольным существительным следует негативная цепочка, например, *прекращение военных действий*, то в целом новая цепочка получит позитив. Поэтому отглагольные существительные выделялись в два отдельных класса: меняющие тональность зависящего от них слова (*nANeg - action negative*, например: *прекращение* или *спад*) и сохраняющие её (*nAPos - action positive*, например: *рост* или *проведение*).

3. Глагольные лексемы и коллокации

Тональная разметка глаголов особенно важна, так как именно глагол определяет, как вычисляется тональность предложения на основе тональности его частей. Тональный словарь глаголов состоит из одиннадцати классов:

- 1 и 2 класс — негативные и позитивные глаголы, определяющие тональность объекта в зависимости от окружения, но независимо от его роли – *vNeg & vPos* – (негативные *уносить*, *стереть*, *освободить от*; позитивные *защищать*, *предрекать*, *болеть за*);
- 3 и 4 класс — негативные и позитивные глаголы, определяющие тональность объекта независимо от окружения, но в зависимости от его роли – *vOppNeg & vOppPos* – (например, глаголы *сдаться* и *проиграть* приписывают негатив субъекту и позитив объекту, а глаголы *обуздать* и *повергнуть*, наоборот, приписывают позитив субъекту и негатив объекту);
- 5 и 6 класс — негативные и позитивные глаголы, определяющие тональность объекта в зависимости от его окружения и роли – *vFlxNeg & vFlxPos* – (в основном в эти классы вошли возвратные глаголы; примеры негативных глаголов: *жаловаться, испугаться, замерзать*; позитивных глаголов: *окупаться*, *влечь*, *согреться*);
- 7 и 8 класс — негативные и позитивные глаголы, определяющие тональность объекта вне зависимости от его роли и окружения – *vNegP & vPosP* – d(например, *расследовать* и *улучшать* всегда приписывают позитив, а *грабить* и *злоупотреблять* — негатив);
- 9 класс — глаголы, соединяющие или приравнивающие тональность объекта и субъекта – *vLink* - (так называемые связочные глаголы: *являться*, *олицетворять*, *относиться*).
- 10 и 11 классы — позитивные и негативные глагольные коллокации – *collvPos & collvNeg*. Например, негативные *наложить руки*, *освободить от должности*, *пробивать насквозь* и позитивные *поразить противника*, *заострить внимание*, *заливаться смехом*.

Списки глаголов составляются с учетом глагольного управления: глаголы, тональность которых менялась в зависимости от глагольного управления, попадали в разные классы (например, *высказаться за* и *высказаться против*).

У всех глаголов и глагольных коллокаций определена сила тональности. Она показывает на сколько сильна эмоциональная оценка у той или иной лексемы или фразы. Сила тональности имеет шкалу от 0 до 2. Например, глаголы из класса *vLink* имеют 0-ю степень экспрессивности, а глаголы и класса *vNegP & vPosP* , как правило, от 1 до 2. С помощью силы тональности определяется как суммарная позитивность / негативность предложения, так и конечная оценка тональности для всего текста.

4. Пополнение словарей

Пред добавлением слова в словарь слово или коллокацию необходимо нормализовать, т.е. привести к единственному числу и именительному падежу. Например, коллокация *курс доллара вырос* будет записана *как курс доллар вырасти*.

Прилагательные не нормализуются, а их окончания заменяются звездочкой «*». Например, прилагательные *красивый/ красивого/ красивая/ красивому* и пр. будут записаны как *красив**

Одно и то же слово может находится в разных классах (т.е. принадлежать разным частям речи), но не может повторяться в одном классе дважды. Поэтому необходимо убедиться, что добавляемое слово отсутствует в соответствующем классе словарей.

Следует добавлять слова в строчку, через перенос строки. Для глаголов и глагольных коллокаций (словари list_collocation_vneg.txt, list_collocation_vpos.txt) через табуляцию требуется также указать целое число от 0 до 2 (оценка силы эмоциональной нагрузки вводимого предиката).

Приложение 2

Пример SQL запроса на удаление неправильно выделенных ключевых тем в БД Semantic

Элемент <список_удаляемых_тем> должен содержать список лексем, соответствующих удаляемым темам и разделенных запятой.

Для параметра table_name указывается название таблицы информационного фонда. При необходимости запрос повторяется для каждого информационного фонда, таблицы которого содержатся в данной БД Semantic.

```
declare @s varchar(1000)
set @s = '<список_удаляемых_тем>'

declare @tmplList table (t nvarchar(100))
insert into @tmplList
  select item from dbo.fn_parse_list(',', @s)

declare @tID table (id int)
     insert into @tID
       select id from dbo.<table_name>_Theme tl, @tmplList tpl
   where tl.ThemeName like tpl.t

delete dbo.<table_name>_ConnDoc where ThemeSourceID in (select
id from @tID)
delete dbo.<table_name>_ConnDoc where ThemeDestID in (select id
from @tID)
delete dbo.<table_name>_ThemeDoc where ThemeID in (select id
from @tID)
print 'dbo.<table_name>_ThemeDoc deleted'
delete dbo.<table_name>_Theme where id in (select id from @tID)
print 'dbo.<table_name>_Theme deleted'
```

Лист регистрации изменений

Изм.	Номера листов (страниц)				Всего листов (страниц) в докум.	№ документа	Входящий № сопроводительного докум. и дата	Подп.	Дата
	измененных	замененных	новых	аннулированных					