IBM Content
Manager for Multiplatforms

Управление системой Information Integrator for Content

Версия 8 Выпуск 2

SH43-0216-01

IBM Content
Manager for Multiplatforms

Управление системой Information Integrator for Content

Версия 8 Выпуск 2

SH43-0216-01

Примечание

Перед тем как использовать данный документ и продукты, описанные в нем, прочтите сведения под заголовком “Замечания” на стр. 133.

Второе издание (март 2003)

Этим изданием можно пользоваться при работе с IBM Enterprise Information Portal for Multiplatforms Версия 8 Выпуск 2 (номер продукта 5724-B43) и со всеми последующими выпусками и модификациями, пока в новых изданиях не будет иных указаний.

Содержание

Об этом руководстве

В этом руководстве излагаются все основные понятия, которые необходимо знать для успешной работы с системой Enterprise Information Portal (EIP). Поскольку EIP содержит несколько компонентов, которыми можно управлять из клиента администратора, и поскольку при помощи EIP можно управлять возможностями других продуктов, эта книга - не типичное руководство по управлению системой. Прежде всего, в ней объясняется, как:

- Использовать EIP для нужд вашего бизнеса
- Обращаться к клиенту администратора системы и использовать его
- Управлять доступом пользователей
- С помощью EIP можно вести поиск информации на нескольких контент-серверах, включая структурированные данные, хранящиеся в реляционных базах данных, неструктурированную или мультимедийную информацию или текстовые документы
- Разрабатывать и реализовывать рабочие потоки и управлять ими

Для кого предназначено это руководство

Эта книга помогает администраторам EIP выполнять следующие задачи:

Управление системой
, Включая управление базой данных, сервером и сетью

Управление пользователями
Определение и предоставление доступа отдельным пользователям и группам, поддержка списков управления доступом

Объединенный поиск
Определение и использование шаблонов объединенного поиска для получения содержимого систем Content Management

Исследование информации
Извлечение информации из документов, категоризация документов и результатов поиска

Поиск в Интернете
Использование программы IBM Web Crawler для поиска и импорта данных из Интернета

Текстовый поиск
Использование IBM DB2 TIE или механизма текстового поиска IBM (только Content Manager Версии 7.1 или более ранней) для поиска и индексирования документов

Поиск изображений
Использование Content Manager Версии 7.1 (или более ранней) для поиска изображений

Управление рабочим потоком
Управление рабочими потоками информации в организации с помощью инструментов рабочего потока EIP

Необходимые для администраторов знания

В зависимости от выполняемой задачи необходимо знать:

- Протоколы защиты для доступа пользователей
- Операционные системы Windows NT, Windows XP, Windows 2000, AIX или Solaris
- Управление сетями
- Модели данных контент-серверов в вашей системе Content Management
- Управление базами данных
- Как применять знания о данных и критериях поиска для создания шаблонов поиска
- Методы и инструменты анализа информации
- Принципы разработки рабочих потоков
- Процессы обработки, которые нужно поддерживать с помощью рабочих потоков EIP

Необходимые для анализа деловой информации или моделирования процессов обработки знания

Специалисты по анализу деловой информации и моделированию процессов обработки найдут в этом руководстве определения основных понятий, связанных с заданием и моделированием рабочих потоков EIP для своей организации.

Чтобы работать с построителем рабочих потоков Enterprise Information Portal, надо:

- Знать требования к персоналу, программы и структуры данных, используемые при обработке деловой информации в вашей организации.
- Уметь принимать решения, касающиеся процессов обработки или рабочих потоков на вашем предприятии.

Где найти дополнительную информацию

Пакет продукта содержит полный комплект информации по планированию, установке, использованию системы и управлению ей. Кроме того, документацию по продукту и поддержку можно получить в World Wide Web.

Информация, включенная в пакет продукта

В пакет продукта включен Информационный центр и все публикации в формате .PDF (Portable document format - формат переносимых документов).

Информационный центр

В пакет продукта включен Информационный центр, который можно установить при установке продукта. Сведения об установке Информационного центра смотрите в разделе *Планирование и установка вашей системы Content Management*.

Информационный центр содержит документацию по Content Manager, Enterprise Information Portal и IBM Content Manager VideoCharger for Multiplatforms. Информация разбита на темы и организована по продуктам и задачам (например, Управление). Кроме механизма навигации и указателей, предусмотрена и возможность поиска.

Публикации в формате PDF

Файлы PDF можно просмотреть с помощью прилагаемой программы Adobe Acrobat Reader для вашей операционной системы. Если у вас не установлена программа Acrobat Reader, ее можно получить на сайте Adobe по адресу www.adobe.com.

В разделе Табл. 1 приводятся публикации по Content Manager, прилагаемые к IBM Content Manager for Multiplatforms.

Таблица 1. Публикации по Content Manager

Имя файла	Заголовок	Номер публикации
install	*Планирование и установка вашей системы Content Management*[1]	GH43-0212-01
migrate	*Перенастройка в Content Manager Версии 8*	SC43-0241-01
sysadmin	*Руководство по управлению системой*	SH43-0213-01

Заказав IBM Content Manager for Multiplatforms, вы получите и IBM Enterprise Information Portal for Mulitplatforms. IBM Enterprise Information Portal for Mulitplatforms можно заказать и отдельно. В Табл. 2 перечислены публикации по Enterprise Information Portal, прилагаемые к продукту.

Таблица 2. Публикации по Enterprise Information Portal

Имя файла	Заголовок	Номер публикации
apgwork	*Application Programming Guide for Windows*[1]	SC27-1347-01
ecliinst	*eClient. Установка, конфигурирование и управление*	SH43-0219-02
eipinst	*Планирование и установка IBM Information Integrator for Content*	GH43-0215-01
eipmanag	*Управление системой EIBM Information Integrator for Content*	SH43-0216-01
messcode	*Сообщения и коды*[2]	SH43-0218-01

Примечание:

1. В *Application Programming Guide for Windows* приводится информация о создании программ и для Content Manager, и для Enterprise Information Portal.
2. В книге *Сообщения и коды* приводятся сообщения и коды для Content Manager и Enterprise Information Portal.

Поддержка в Web

Поддержка данного продукта доступна в Web. Для этого выберите опцию **Support** (Поддержка) на Web-сайтах продукта по адресам:

www.ibm.com/software/data/cm/

www.ibm.com/software/data/eip/

Документация поставляется в электронной форме вместе с продуктом. Если вам понадобится получить доступ к документации по продукту, имеющейся в Web, перейдите на Web-сайт продукта и щелкните по опции **Library** (Библиотека).

В WWW также имеется интерфейс для работы с документацией в формате HTML, который называется Enterprise Documentation Online (EDO). В настоящее время он содержит справочную информацию по API. Информацию о том, как получить доступ к EDO, смотрите на Web-странице Enterprise Information Portal Library.

Как послать ваши отзывы

Обратная связь поможет фирме IBM поставлять качественную информацию. Пожалуйста, высылайте любые свои замечания об этой книге или какой-либо другой документации по Content Manager или Enterprise Information Portal. Выслать замечания можно:

- Через Web. Зайдите на страницу IBM Data Management Online Reader's Comment Form (RCF) по адресу:

  www.ibm.com/software/data/rcf

  Эту страницу можно использовать для ввода и отправки замечаний.
- По электронной почте на адрес comments@vnet.ibm.com. Не забудьте указать название продукта, номер версии продукта, название и номер книги (если есть). Если вы шлете замечание к определенному тексту, укажите положение этого текста (например, главу и название раздела, номер таблицы, номер страницы или заголовок темы справки.)

Что нового в EIP Версии 8.2

В продукт внесены следующие изменения:

Поддержка Sun Solaris
В системе Solaris можно установить соединители, возможности и базы данных.

Общее управление системой
Одна прикладная программа клиента обеспечивает независимый доступ к управлению Content Manager и Enterprise Information Portal.

Новые соединители

- Соединитель ICM для Content Manager Версии 8 Выпуск 1 позволяет использовать преимущества мощной системы хранения документов Content Manager Версии 8.
- Новый соединитель C++ Extended Search Версии 3.7 работает в AIX.

Улучшенные соединители

- Параметрический текстовый поиск поддерживается и на уровне объединения, и при прямом соединении Extended Search.
- В соединитель OnDemand внесены функциональные усовершенствования и улучшения производительности , в том числе:
  - Изменения в структуре DDO OnDemand.
  - Поддерживается асинхронный поиск

Новые службы исследования информации

- Извлечение признаков
- Кластеризация
- Идентификация языка

IBM Web Crawler
IBM Web Crawler - возможность, которая позволяет пользователям искать информацию в Web и базах данных Lotus Notes и составлять для нее сводки.

Усовершенствования рабочих потоков
Рабочий поток теперь полностью поддерживается в AIX и Solaris. Построитель рабочего потока, API и JavaBeans обеспечивают усовершенствованные функции рабочего потока и удобство использования.

Информационный центр

Информационный центр, доступный через браузер, содержит документацию по Content Manager, Enterprise Information Portal и IBM Content Manager VideoCharger for Multiplatforms. Информация разбита на темы и организована по продуктам и задачам (например, Управление). Кроме механизма навигации и указателей, предусмотрена и возможность поиска.

Доступность

Функции доступности помогают пользователю с физическими недостатками, например с ограниченной подвижностью или недостаточным зрением, с успехом пользоваться программными продуктами. Основные функции доступности продукта:

- Возможность использовать клавиатуру вместо мыши для работы с любыми функциями.
- Поддержка улучшенных свойств дисплея.
- Опции зрительных и звуковых оповещений.
- Совместимость с технологиями для людей с физическими недостатками
- Совместимости с возможностями доступности операционной системы
- Удобные форматы документации

Глава 1. Введение в Enterprise Information Portal

Многим предприятиям, например, страховым компаниям и финансовым учреждениям, приходится иметь дело с очень большим объемом деловых документов. Средства управления и доступа к деловой информации требуются предприятиям в самых разных сферах бизнеса.

Контент-сервер - это программная система, обеспечивающая хранение мультимедийных данных, деловых форм, документов и связанных с ними данных и метаданных и позволяющая сотрудникам работать с этим содержанием. Если не существует способа эффективно соединяться с разнородными контент-серверами, предприятие может впустую тратить время и деньги на дублирование информации или же приучить сотрудников выполнять поиск многократно.

Система Enterprise Information Portal (EIP) представляет передовую технологию, которая делает доступными все информационные ресурсы предприятия с вашей рабочей станции. Соединяя клиент с разнородными серверами, EIP помогает добиться максимальной доступности информации и активов мультимедиа. С помощью EIP пользователи клиентов могут быстро и одновременно обращаться ко всем подключенным контент-серверам. Пользователи могут также выполнять анализ информации или сложный поиск в данных контент-серверов (включая серверы внутрикорпоративной сети и серверы в Интернете). Они могут выполнять задачи рабочих потоков, входящие в определенные вами процессы обработки информации предприятия.

С помощью EIP можно настроить прикладные программы для вашего предприятия. Используя примеры EIP, прикладные программисты могут создавать программы и для настольных систем, и для Web.

В этом разделе приводится обзор EIP. Возможности и функции EIP показаны на примере работы вымышленной страховой компании XYZ Insurance.

Поиск информации о клиентах

XYZ Insurance (XYZ) - большая компания по страхованию имущества от несчастных случаев, она хранит много фотографий, страховых исков, полисов, актов оценки страхового имущества, докладов экспертов и других документов.

Компания XYZ хранит все письма держателям полисов, медицинские формы и формами оценки в цифровом виде в файловых картотеках Lotus Domino.Doc. Все условия полисов, уведомления и счета компания XYZ архивирует на сервере Content Manager OnDemand для длительного хранения и быстрого доступа к ним. Все формы страховых исков, фотографии и письма, полученные от держателей полисов, XYZ хранит в папке системы Content Manager for iSeries. Доклады экспертов XYZ хранит в менеджере каталогов данных Data Warehouse Center DB2 Universal Database (DB2 UDB) Кроме того, у компании XYZ есть мультимедийные активы, например, графика высокого разрешения, для отделов рекламы, связей с общественностью и развития, которые AA хранит в системе Content Manager. Информацию о процедурах действий компании XYZ хранит во внутренней сети компании.

Потребности

Для обработки страховых исков, ответов на звонки клиентов и общего обслуживания держателей полисов недостаточно содержания только одного сервера, так как сотрудникам необходим доступ ко всей информации о клиентах. Для обслуживания клиентов сотрудникам требуется одновременный доступ к разнообразным контент-серверам. Компании XYZ необходимо решение, которое позволит объединить различные контент-серверы и внутреннюю сеть для поиска и получения информации. Требуется также расширить использование рабочих потоков.

Доступ к документам нужен многим сотрудникам, в том числе клеркам, специалистам по искам и страховым агентам. Компания XYZ должна ограничить доступ к одним элементам, предоставив неограниченный доступ к другим. Кроме того, компании XYZ требуется простой в использовании интерфейс, не требующий много времени на освоение.

Решение

Компания XYZ Insurance использует EIP, поскольку его мощные технологии поиска позволяют выполнять поиск данных на всех контент-серверах. Теперь когда в Центр обработки запросов компании XYZ поступает запрос, для получения всей необходимой информации о держателе страхового полиса достаточно одной операции объединенного поиска.

XYZ Insurance использует также возможность исследования информации EIP для поиска и получения информации из внутренней сети компании. Она также хочет расширить использование рабочих процессов.

Обзор

EIP - это комплексный продукт, его компоненты работают совместно и обеспечивают решение, которое идеально соответствует потребностям вашего предприятия. На базе многоуровневой архитектуры EIP предоставляет клиент управления для управления поисками, клиенты для выполнения поиска и соединители для соединения с разнородными контент-серверами, такими как IBM Content Manager, Content Manager ImagePlus for OS/390, Content Manager OnDemand, Lotus Domino.Doc, DB2 Universal Database, DB2 DataJoiner и Менеджер каталогов данных Центра хранилищ данных DB2. Для других контент-серверов можно при помощи комплекта и примеров соединителей EIP написать дополнительные соединители.

Архитектура EIP позволяет клиентским программам в одной операции осуществлять поиск сразу по нескольким контент-серверам. Чтобы осуществить поиск, клиент использует шаблоны поиска, определенные администратором EIP.

Используя шаблоны поиска, клиент выполняет *объединенный поиск* - поиск, который выполняется одновременно на нескольких контент-серверах; собственные атрибуты этих серверов были отображаются на атрибуты объединения, используемые в шаблоне поиска. Шаблоны поиска EIP содержат критерии поиска, использующие атрибуты объединения, которые отображаются на собственные атрибуты каждого контент-сервера. Шаблоны поиска создает администратор EIP. EIP содержит соединители для поиска данных на многих контент-серверах и обращения к ним. Затем контент-серверы возвращают объекты данных клиенту.

Архитектура EIP обеспечивает следующие преимущества:

- Обращение по одному запросу сразу к нескольким различным контент-серверам, которые поддерживают транзакции e-business и программы обслуживания пользователей.

- Возможность исследование информации сразу с нескольких контент-серверов, в том числе и из Web.
- Доступ процесса рабочего потока к данным на нескольких различных контент-серверах.
- Поддержка разработки клиентских программ, не зависящих от расположения данных на определенных контент-серверах, благодаря разделению клиентских программ, индексов и данных.

Введение в компоненты Enterprise Information Portal

В этом разделе описаны все компоненты EIP и опции установки.

В Табл. 3 на стр. 3 перечислены компоненты и поддерживаемые операционные системы.

Таблица 3. Совместимость компонентов EIP с операционными системами

Компонент	Windows	AIX	Solaris	Примечания
Управляющая база данных	да	да	да	База данных включает в себя функции построителя рабочих потоков
Клиент администратора	да	нет	нет	Клиент может соединяться с базами данных, установленными в операционных системах Windows, AIX или Solaris.
Соединители	да	да	да	
Исследование информации	да	да	да	
IBM Web Crawler	да	да	да	
Клиент текст. поиска	да	да	да	
Клиент поиска изображений	да	да	да	
Комплект и примеры соединителей	да	да	да	• Версия Windows содержит исходный код для компиляции примера клиента. В AIX исходный код клиента не устанавливается. • Примеры рабочего потока и API установлены с примером соединителя объединения.

Таблица 3. Совместимость компонентов EIP с операционными системами (продолжение)

Компонент	Windows	AIX	Solaris	Примечания
Программа просмотра	да	нет	нет	Устанавливает клиент и программу просмотра OnDemand.
Информационный центр	да	да	да	

Управление

Компонент управления содержит управляющую базу данных и клиент администратора. При установке управляющей базы данных вы устанавливаете также возможность рабочих потоков.

Управляющая база данных: Управляющая база данных - это база данных DB2, где хранится информация о пользователях и группах EIP, уровнях привилегий, паролях, ID пользователей и другая информация. Эта база данных поддерживает также возможности рабочих потоков и (необязательно) исследования информации. Можно установить несколько баз данных. Каждая база данных поддерживает возможности рабочих потоков EIP. Если у вас установлена система Content Manager Версии 8, базу данных библиотечного сервера Content Manager Версии 8 можно использовать и в качестве управляющей базы данных EIP. База данных библиотечного сервера содержит всю информацию, требуемую для EIP.

Клиент администратора: Клиент администратора может установить только на рабочие станции Windows. Можно установить несколько таких клиентов. Если у вас установлена система Content Manager Версии 8, вы можете управлять EIP и Content Manager Версии 8 с одного клиента.

Интерфейс этого клиента позволяет администратору:

- Определять контент-серверы для объединенного поиска.
- Указывать собственные элементы и атрибуты контент-серверов и их отображение на элементы объединения.
- Поддерживать перечень собственных объектов и атрибутов для всех пользовательских контент-серверов.
- Создавать шаблоны поиска.
- Определять пользователей и группы и управлять ими.
- Назначать пользователям привилегии и наборы привилегий.
- Определять доступ к шаблонам поиска и задавать условия, определяющие, что может делать пользователь с полученной в результате поиска информацией.
- Создавать процессы рабочих потоков и управлять ими.

Соединители

Соединители обеспечивают интерфейс между клиентами EIP, контент-серверами и управляющей базой данных. Соединители контент-серверов, такие как соединитель Content Manager Версии 7.1, поддерживают функции, которые позволяют EIP регистрироваться на сервере, искать информацию и возвращать ее клиенту администратора или клиентам конечных пользователей. Соединитель объединения связывает клиент администратора с управляющей базой данных.

EIP предлагает следующие соединители:

- Соединитель объединения, который связывает клиент EIP с управляющей базой данных.
- Соединитель реляционных баз данных для DB2 Universal Database 7.1, драйвер JDBC 1.3 (только для Java), ODBC 3.0 (только для C++), DataJoiner 2.1.1.
- Соединитель Content Manager для серверов Content Manager Версии 7.1.
- Соединитель Content Manager для серверов Content Manager Версии 8.2.
- Соединитель Content Manager OnDemand для Content Manager OnDemand Версии 7.1.
- Content Manager for VisualInfo for 400 Версии 4.3 и Версии 5.1.
- Соединитель Content Manager ImagePlus for OS/390 для ImagePlus/390 Folder Application Facility Версии 3.1, Image Plus/390 ODM Версии 3.1.
- Соединитель Lotus Domino.Doc для Domino.Doc Версии 3.0a, Desktop Enabler Версии 3.0a.
- Соединитель Extended Search для Версии 3.7.
- Соединитель менеджера каталогов данных для DB2. Universal Database Visual Warehouse Версии 5.2, DB2 Universal Database Версии 7.2.

Возможности

У EIP есть четыре дополнительные возможности.

Исследование информации

Компонент Исследование информации поддерживает лингвистические службы для поиска скрытой информации в текстовых документах на контент-серверах. При обработке текстовых документов создаются метаданные, которые можно использовать для составления сводок, категоризации и поиска. Для исследования информации требуется сервер прикладных программ WebSphere 4.0 (стандартное или расширенное издание). Кроме того, вы можете сгруппировать похожие документы, извлекать из документов данные, например, фамилии или названия фирм, и определять язык документа.

Клиент поиска изображений

Обеспечивает интерфейс для вызова поиска изображений и управления им на контент-серверах Content Manager Версии 7.

Клиент текстового поиска

Обеспечивает интерфейс для вызова текстового поиска и управления им на сервере текстового поиска.

IBM Web Crawler

Web Crawler - это написанный на языке Java инструмент для поиска и анализа содержимого. Web Crawler может работать с содержимым в интрасети, в объединении таких сетей, в Интернете, в базах данных Lotus Notes (напрямую или через Domino), в локальных файловых системах и через соединения FTP.

Web Crawler может исследовать метаданные и текстовые данные для различных типов содержимого. Например, содержимое HTML может анализироваться по URL, заголовку, телу документа, времени последнего изменения и метатегам, таким как автор, ключевые слова, описание и т.д. Пользователь выбирает нужный вариант анализа из заранее предопределенных для данного типа содержимого. Содержимое и/или анализируемые метаданные сохраняются на локальном диске.

Программа просмотра содержимого

При установке программы просмотра OnDemand устанавливается клиент OnDemand и другие файлы, необходимые для просмотра документов, полученных с сервера OnDemand.

Комплекты и примеры соединителей

В EIP входит комплект соединителей, включающий примеры программ, которые можно использовать для проверки различных функций EIP, например:

- соединения с контент-серверами и отсоединения от них
- выполнения операторов SQL и других примеров запросов на контент-серверах
- определения типов MIME контент-серверов и т.д.

Комплект соединителей для Windows: Чтобы установить комплект и примеры соединителей на серверах Windows, нужно выбрать тип компьютера Рабочая станция разработки. Затем выбрать компонент Комплект и примеры соединителей. Примеры программ можно установить для всех соединителей или выбрать отдельные примеры для установленных соединителей.

На серверах Windows примеры программ комплекта соединителей организованы следующим образом:

`c:\CMBROOT\SAMPLES\activex\`*xx*

`c:\CMBROOT\SAMPLES\cpp\`*xx*

`c:\CMBROOT\SAMPLES\java\`*xx*

`c:\CMBROOT\SAMPLES\jsp\`*xx*

`c:\CMBROOT\SAMPLES\server\`*xx*

где *xx* - имя каталога, содержащего примеры программ для каждого применимого соединителя (например, `db2`, `od`, `dl` и т.д.).

На серверах AIX примеры программ организованы следующим образом:

`/usr/lpp/cmb/samples/cpp/icm`

`/usr/lpp/cmb/samples/java/`*xx*

`/usr/lpp/cmb/samples/jsp/`*xx*

`/usr/lpp/cmb/samples/server/exit`

где *xx* - имя подкаталога (например, `beans`, `servlets` и т.д.).

На серверах Solaris примеры программ организованы следующим образом:

`/opt/IBMcmb/samples/java/`*xx*

`/opt/IBMcmb/samples/jsp/`*xx*

`/opt/IBMcmb/samples/server/exit`

где *xx* - имя подкаталога (например, `beans`, `servlets` и т.д.).

Примеры программ включают документацию, в которой описываются программы и указываются параметры сервера (параметры среды, памяти и т.д.), необходимые для работы примеров программ.

Информационный центр

Компонент Информационный центр содержит информационный центр Enterprise Information Portal. Информационный центр - это версия библиотеки Enterprise Information Portal, содержащая документы в формате Web и допускающая поиск.

Глава 2. Клиент администратора. Введение

Клиент администратора - это интерфейс между управляющей базой данных EIP и администратором EIP. В этом разделе описываются многие из возможностей и функций этого клиента, помогающие управлять системой EIP.

Для вызова некоторых возможностей и функций, таких как определения сервера и управление пользователями, используются значки в левой панели окна клиента. Другие функции вызываются из панели инструментов.

Использование Первых шагов для ознакомления с клиентом администратора

Первые шаги - это модуль, поставляемый с каждой установкой EIP. Он содержит данные примера и использует их для заполнения объектов, поэтому вам не обязательно использовать реальные данные. Используйте *Первые шаги*, чтобы научиться работать с определениями серверов, пользователями и группами и другими возможностями, понять основную структуру клиента администратора и ознакомиться с ним.

Управление EIP

Как системный администратор, вы можете выполнять в клиенте администратора следующие задачи:

- Определение контент-серверов
- Управление пользователями и группами
- Управление привилегиями и уровнями доступа
- Создание шаблонов объединенного поиска
- Создание объектов объединения
- Создание поддоменов, если разрешено создание доменов администраторов.
- Работа с рабочим потоком, если эта возможность включена.
- Создание текстового объекта объединения в Content Manager Версии 7.

Управление пользователями и группами

Вы управляете доступом пользователей к поиску документов на нескольких контент-серверах и работе с ними, создавая ID пользователей и назначая им привилегии. Для ограничения доступа к данным, хранящимся в системе, вы определяете и назначаете пользователям соответствующие привилегии.

Использование инструментов клиента администратора

В этом разделе описываются инструменты, входящие в клиент администратора.

Конфигурация LDAP

Когда вы выберете эту опцию, EIP откроет окно с четырьмя вкладками:

- Вкладка LDAP - здесь вы можете разрешить импорт источников данных с сервера LDAP, импорт и аутентификацию пользователей LDAP или и то, и другое.

- Вкладка Сервер - содержит поля для задания спецификаций сервера LDAP, включая имя хоста, имя пользователя, использование ссылок на другие серверы и так далее.
- Вкладка Аутентификация - содержит поля для задания информации защиты SSL (Secure Sockets Layer).
- Вкладка Дополнительные - на ней задаются максимальное число записей и срок ожидания сервера.

Опция отображения пользователей
Эта опция позволяет отключить используемое по умолчанию отображение пользователей.

Редактор отображения пользователей объединения
Редактор пользователей объединения показывает список пользователей и позволяет вам отображать пользователей для определенных контент-серверов.

Программа просмотра шаблонов поиска
Программа просмотра шаблонов поиска показывает подробную информацию обо всех шаблонах поиска. В ней есть три опции просмотра подробностей шаблонов поиска:

- Связанные отображения (опция по умолчанию) - показывает информацию об объектах объединения и другую информацию о шаблоне поиска
- Шаблон поиска - показывает информацию об операции по умолчанию, значениях по умолчанию и т.п.
- Вывод результатов - показывает информацию об имени окна вывода, его ширине, порядке критериев и т.п.

Программа просмотра перечней сервера
Показывает перечни для выбранного сервера или серверов.

Программа просмотра журнала
Окно журнала служит для просмотра журнала, генерируемого после обновления перечня сервера. Этот журнал содержит список сообщений о расхождениях между новым и предыдущим перечнями.

Службы
Выберите службы, чтобы разрешить рабочий поток и/или исследование информации.

Домены администраторов
Выберите домены администраторов, чтобы включить их. Включив однажды домены администраторов, вы не сможете отключить их позже.

Редактор типов MIME
Редактор типов MIME выводит следующую информацию для каждого контент-сервера:

- Класс содержимого
- Расширение файла
- Столбец реляционной базы данных (RDB)
- Тип MIME

В редакторе типов MIME имена контент-серверов выводятся в виде аббревиатур и соответствуют именам в списке имен контент-серверов, показываемых при определении нового контент-сервера. **Подсказка**: DL - аббревиатура для контент-сервера Content Manager Версии 7.1. V4 - аббревиатура для контент-сервера Content Manager for AS/400.

В редакторе типов MIME можно удалить и изменить информацию по умолчанию или добавить новую информацию.

Редактор связей MIME с прикладными программами
В редактор связей MIME с прикладными программами можно удалить или изменить пять заданных по умолчанию связей MIME с прикладными программами или добавить новые связи. Значения и параметры, заданные в редакторе связей MIME с прикладными программами, определяют программы просмотра, используемые клиентами конечных пользователей.

Определение типа сервера
Этот инструмент служит для определения всех пользовательских серверов, разработанных вашими системными программистами.

Изменить ID/пароль DB2
Выберите эту опцию, чтобы изменить ID пользователя или пароль DB2 только для соединения. Этот ID пользователя совершенно самостоятелен по отношению к ID пользователя администратора.

О привилегиях

В этом разделе описываются привилегии Enterprise Information Portal. Разверните значок Авторизация, чтобы получить доступ к следующим четырем привилегиям. **Примечание:** поскольку управление Content Manager Версии 8 и EIP Версии 8 возможно с одного и того же клиента, клиент показывает все привилегии для обеих систем.

Привилегии

Клиент администратора содержит предопределенные привилегии, группы привилегий и наборы привилегий. Привилегии дают пользователям системы (таким, как администраторы или конечные пользователи клиента) право или права выполнять определенные действия с определенными объектами.

Привилегии
EIP поддерживает различные привилегии. Привилегия - это право выполнять определенное действие с определенным объектом. Например, конечным пользователям клиента можно предоставить привилегии ItemAdd и ItemDelete, дающие право добавлять элементы на контент-сервер и право удалять элементы с него. Чтобы просмотреть привилегии, разверните значок Авторизация и дважды щелкните по Привилегии. Чтобы создать привилегии, щелкните правой кнопкой мыши по Привилегии и выберите Создать.

Группы привилегий
В EIP есть группы привилегий по умолчанию. Группа привилегий - это собрание связанных привилегий. Например, группа привилегий Управление EIP содержит пять привилегий, обычно связанных с управлением системой EIP:

- EIPAdminServer
- EIPAdminEntity
- EIPAdminTextEntity
- EIPAdminTemplate
- EIPAdminInfoMining

Чтобы просмотреть или изменить привилегии, назначенные группе привилегий, разверните значок Авторизация и дважды щелкните по имени группы привилегий. Чтобы создать группы привилегий, щелкните правой кнопкой мыши по Группы привилегий и выберите Создать.

Наборы привилегий

В EIP есть несколько наборов привилегий по умолчанию. Наборы привилегий - это собрания привилегий, определяющих роли пользователей. Например, набор привилегий ClientUserCreateAndDelete содержит 17 привилегий, связанных с ролями конечных пользователей - например, ItemDelete (удаление элемента), ItemAdd (добавление элемента) и т.д. Если при создании ID пользователя для пользователя клиента ему назначен набор привилегий ClientUserCreateAndDelete, этот пользователь может зарегистрироваться на контент-сервере и выполнять любые из этих 17 действий пользователя, входящих в этот набор привилегий. Чтобы просмотреть или изменить набор привилегий, разверните значок Авторизация, щелкните по Наборы привилегий и дважды щелкните по имени набора привилегий. Чтобы создать наборы привилегий, щелкните правой кнопкой мыши по Наборы привилегий и выберите Создать.

Управляя пользователями и группами, вы связываете набор привилегий с пользователем и/или группой пользователей. Если наборы привилегий назначены группе пользователей, каждый пользователь в этой группе может выполнять все действия, входящие в эти наборы привилегий.

Переключение окон продуктов и баз данных

Если ваша организация работает и с Content Manager, и с Enterprise Information Portal, у вас есть доступ к обоим клиентам администратора системы через один пользовательский интерфейс. Ранее, если у вас были установлены оба продукта, вам приходилось открывать два отдельных клиента. Переключение с одного клиента на другой - удобный путь изменения информации, относящейся к обоим клиентам, и быстрого доступа к обоим продуктам.

Чтобы переключиться с управления EIP на управление Content Manager, не выходя из системы, выберите Content Manager в выпадающем меню над левой панелью в главном окне управления системой.

Чтобы переключиться с одной базы данных объединения на другую, дважды щелкните по значку базы данных объединения на левой панели окна клиента.

У вас есть также возможность управлять разными базами данных, не выходя из клиента и не регистрируясь в новой базе данных. Для каждой управляющей базы данных, перечисленной в файле `cmbds.ini`, клиент администратора выводит специальный значок. Чтобы переключиться на другую базу данных, щелкните по этому значку. Если для новой базы данных нужен другой ID пользователя по отношению к введенному вами при регистрации на клиенте, система попросит вас ввести другой ID пользователя.

Улучшения и усовершенствования клиента администратора

В EIP Версии 8.2 в клиент администратора EIP внесены существенные усовершенствования, в том числе:

Улучшенные мастера и диалоги

Новые диалоги облегчают управление пользователями. Новые мастера облегчают определение и изменение объектов объединения и шаблонов поиска. Пользователи по-прежнему могут выбрать использование диалоговых окон, поддерживавшихся EIP Версии 7.1.

Общий клиент администратора

При установке EIP Версии 8.2 и Content Manager Версии 8.2 в одной системе

эти два продукта используют общий клиент администратора. Если вы - администратор обоих этих продуктов, вы регистрируетесь только один раз и переключаетесь между двумя программами внутри клиента. Можно также переключаться между управляющими базами данных, не выходя каждый раз из системы и не регистрируясь заново.

Администраторы доменов
Можно создавать администраторов доменов с административными полномочиями только для заданного домена.

Поддержка единой регистрации и LDAP
EIP теперь использует Windows Active Directory и LDAP, чтобы разрешить пользователям доступ к нескольким контент-серверам после одной регистрации.

Соединение клиента администратора с локальной управляющей базой данных

Если управляющая база данных установлена на том же сервере, что и клиент администратора, информация, необходимая для соединения локального клиента с сервером, уже хранится в файле `cmbds.ini`, содержащем информацию о соединении с базой данных. После установки не требуется выполнять никаких действий по конфигурированию и можно сразу установить соединение, следуя инструкциям в этом разделе. **Требование:** Если при помощи утилиты установки базы данных EIP создаются дополнительные локальные базы данных, нужно вручную внести необходимую информацию в файл `cmbds.ini` перед соединением с этой новой базой данных.

1. Выберите **Пуск-->Программы-->Enterprise Information Portal for Multiplatforms 8.2-->Управление**
2. Выберите локальную базу данных из выпадающего списка в поле Сервер.
3. Введите ID и пароль пользователя администратора и нажмите кнопку OK.
4. Откроется клиент администратора системы. **Подсказка:** Если вы использовали Первые шаги EIP, на левой панели клиента будут показаны примеры баз данных.

Соединение клиента администратора с удаленной управляющей базой данных

Установить соединение клиента администратора EIP с удаленной базой данных AIX, Windows или Solaris можно двумя способами:

- Соединиться через сервер RMI (дополнительную информацию смотрите в книге *Планирование и установка Information Integrator for Content*).
- Определить соединение, внеся базу данных в каталог при помощи Ассистент конфигурирования DB2, и затем определить параметры соединения с сервером при помощи утилиты конфигурирования сервера EIP. Эта утилита копирует в файл `cmbds.ini` такую информацию, как имя схемы базы данных, алиас, информацию об операционной системе и т.д. При запуске клиента администратора системы он получает список серверов, на которых вы можете регистрироваться, из серверов, определенных в файле `cmbds.ini`.

Требование: Каждую удаленную базу данных нужно внести в каталог по отдельности. Чтобы с удаленной базой данных можно было соединиться из клиента администратора, она должна быть перечислена в файле `cmbds.ini`.

Совет: Если вы - опытный пользователь, вы можете пропустить шаги использования утилиты конфигурирования сервера и изменить файл `cmbds.ini` с помощью текстового редактора. По умолчанию файл `cmbds.ini` находится в каталоге `C:\Program Files\IBM\CMgmt`.

Внимание: Если тот, кто устанавливал этот продукт, уже сконфигурировал значения каталога баз данных для удаленной базы данных, с которой вы хотите соединяться, вам не нужно выполнять шаги для использования Ассистента конфигурирования DB2 для этой базы данных. Но если нужные значения каталога баз данных не заданы или если вы хотите соединяться с дополнительной удаленной базой данных, нужно использовать Ассистент конфигурирования DB2 и внести в файл `cmbds.ini` параметры соединения для этой дополнительной базы данных (или баз данных).

Шаг 1 - Внесение удаленной базы данных в каталог при помощи Ассистента конфигурирования DB2

Ассистент конфигурирования DB2 (CCA) вносит в каталог DB2 удаленную базу данных EIP. Чтобы внести в каталог удаленную базу данных при помощи DB2CCA, нужно знать имя хоста удаленного сервера, имя базы данных и номер порта экземпляра базы данных, а также нужно определить алиас для этой удаленной базы данных.

В шагах 1a - 1f объясняется, как узнать имя базы данных, имя схемы и номер порта соединения. Эти имена и номер порта нужно знать, чтобы задать их для соединения между клиентом администратора и удаленной базой данных.

1. Получите информацию о соединении с удаленной базой данных:
   a. Зарегистрируйтесь на удаленном сервере AIX, Windows или Solaris с ID пользователя с полномочиями администратора DB2.
   b. Введите `db2 list db directory`
   c. Выберите имя управляющей базы данных, с которой хотите соединиться. Запишите, в каком экземпляре DB2 установлена эта база данных (у разных экземпляров могут быть разные номера портов соединения).
   d. Введите `db2 connect to <база_данных> user <ID_пользователя> using <пароль>`
   e. Введите `db2 list tables` и запишите имя схемы базы данных (необходимое для утилиты конфигурирования сервера).
   f. Узнайте номер порта соединения, связанный с этой удаленной управляющей базой данных:

      В Windows:

      1) Откройте Центр управления DB2 на удаленном сервере Windows.
      2) Щелкните правой кнопкой мыши по одному из экземпляров, доступных для локального компьютера.
      3) Выберите ″Задание связи...″.
      4) Нажмите кнопку ″Свойства″ справа от выбора TCP/IP. В окне будет показан номер порта.

      В AIX или Solaris

      1) Введите `cd /usr/etc`
      2) Введите `cat services`
      3) Используя прокрутку, найдите в списке служб номер порта соединения для экземпляра этой удаленной базы данных. Например, если база данных установлена в экземпляре `db2inst1`, порт соединения может быть `50000`.

4)

2. Используйте Ассистент конфигурирования DB2 для внесения этой удаленной базы данных в каталог. Дополнительную информацию смотрите в справке DB2CCA.
   a. Зарегистрируйтесь на сервере Windows, где установлен клиент администратора. Необходимо зарегистрироваться с ID пользователя, обладающим полными привилегиями DB2ADM.
   b. Вызовите Ассистент конфигурирования DB2 из меню Пуск-->Программы.
   c. Следуя указаниям Ассистента конфигурирования DB2, внесите удаленную базу данных в каталог и проверьте соединение с ней.
   d. Если проверка соединения в DB2 CCA была успешной, выполните шаги в разделе “Шаг 2 - Использование утилиты конфигурирования сервера” или напрямую измените файл `cmbds.ini`, задав хранимые в нем параметры соединения с удаленной базой данных.

Шаг 2 - Использование утилиты конфигурирования сервера

Утилита конфигурирования сервера предложит вам ввести информацию о соединении (номер порта, имя хоста и т.п.) с удаленной базой данных и сохранит данные в файле `cmbds.ini`.

1. Выберите **Пуск-->Программы-->IBM Enterprise Information Portal for Multiplatforms-->Утилита конфигурирования сервера**.
2. Введите информацию в поля (смотрите раздел Табл. 4).

Таблица 4. Утилита конфигурирования сервера

Поле	Информация	Примечания
Сервер	Выберите тип базы данных - Content Manager или EIP.	Здесь сервер - это тип базы данных, а не имя сервера, на котором установлена база данных. **Совет:** При помощи клиента администратора можно управлять обоими типами баз данных, только если в вашей системе клиенты администратора Content Manager и EIP находятся на одном компьютере.
Имя сервера	Введите алиас базы данных, с которой вы соединяетесь. Требование: Нужно использовать тот же алиас, что был определен в DB2CCA.	Алиас - это уникальное обозначение удаленной базы данных на вашей рабочей станции. Максимальная длина алиаса - восемь символов. Например, имя базы данных может быть ICMNLSDB, а алиас - REMOTE1.
Имя схемы	Введите имя схемы, назначенной при создании удаленной базы данных.	По умолчанию для баз данных EIP и Content Manager используется имя схемы ICMADMIN.
Имя хоста	Введите имя компьютера, на котором установлена удаленная база данных.	Введите полное имя хоста или IP-адрес компьютера, на котором установлена удаленная база данных.

Таблица 4. Утилита конфигурирования сервера (продолжение)

Поле	Информация	Примечания
Операционная система	Выберите операционную систему из выпадающего списка.	Выберите AIX, Sun Solaris или Windows. Опция OS/390 недоступна в EIP 8.2.
Номер порта	Введите номер порта, назначенный удаленной базе данных.	Номер порта соединения по умолчанию для баз данных EIP и Content Manager, установленных в Windows, AIX и Solaris - 50000.
Имя удаленной базы данных	Введите имя удаленной базы данных. Используйте заглавные буквы.	ICMNLSDB - имя по умолчанию для баз данных EIP и Content Manager.
Имя узла	Введите имя узла удаленной базы данных EIP или Content Manager.	Имя узла - это уникальное имя, назначаемое удаленной базе данных, аналогичное алиасу, который вы создаете для удаленной базы данных. Чтобы найти имя узла базы данных, установленной на сервере Windows, AIX или Solaris: a. Откройте сеанс командной строки db2. b. В ответ на приглашение `db2=>` введите `LIST NODE DIRECTORY` c. DB2 выведет имена узлов и другую информацию для всех баз данных, установленных или определенных на удаленном сервере.
Разрешить единую регистрацию	Включите, если база данных установлена с поддержкой единой регистрации.	Значение по умолчанию - выключено (запрещено).
Опции защиты	Выберите аутентификацию клиентов, если эта возможность была выбрана при создании базы данных.	Значение по умолчанию - сервер.

3. Нажмите кнопку OK.

Шаг 3 - Проверка соединения с удаленной базой данных

1. Зарегистрируйтесь на сервере Windows, где установлен клиент администратора.
2. Выберите **Пуск-->Программы-->Enterprise Information Portal for Multiplatforms 8.2-->Управление**.
3. Выберите алиас удаленной базы данных из выпадающего списка в поле Сервер. Это алиас, который вы определили в утилите конфигурирования сервера и Ассистенте конфигурирования DB2.
4. Введите ID пользователя и пароль для удаленной базы данных.
5. Нажмите кнопку OK. Откроется клиент администратора.

Определение типов документов

EIP позволяет просматривать документы различных типов. Определив тип документа, вы можете открывать эти документы в соответствующих прикладных программах. Например, если вы храните документы на своем сервере OnDemand Content Manager Lotus Word Pro, можно настроить EIP, чтобы он открывал документы с расширением `.lwp` в Lotus Word Pro, а не в клиентской программе просмотра документов.

Чтобы определить тип документа, измените файл `cmbcc2mime.ini` в каталоге CMBROOT (обычно это каталог *x*`:\Program Files\IBM\CMgmt`). Этот файл содержит инструкции по разработке пользовательских определений MIME. В этом файле для классов содержимого определяются соответствующие потоки типа MIME, чтобы клиент мог читать содержимое контент-серверов.

Внимание: При запуске какой-либо прикладной программы для конкретного типа MIME выводится только базовый объект. Разметка в документе не отображается. Если в документе несколько частей, на экран выводится его первая часть. Типы MIME в обоих файлах должны совпадать.

Изменение файла типов MIME для сервера (cmbcc2mime.ini)

При добавлении типов MIME для сервера убедитесь, что добавляемый тип документа является одним из типов MIME, созданным для этого файла. Дополнительную информацию смотрите на сайте:`ftp://ftp.isi.edu/in-notes/iana/assignments/media-types`.

Чтобы добавить значения в файл `cmbcc2mime.ini`:

1. Откройте `cmbcc2mime.ini` в текстовом редакторе.
2. Задавайте свои значения в следующем формате:
   - Класс содержимого начинается с 4096
   - За значением класса содержимого следует знак равенства (=)
   - За знаком равенства должен следовать тип MIME. Если это не стандартный тип MIME для данного класса содержимого:
     a. Тип MIME состоит из типа и подтипа. Допустимые типы: программа, текст, изображение, модель, сообщение, аудио и видео.
     b. За типом следует косая черта (/)
     c. Чтобы создать подтип, перед обозначением класса документа поставьте маркер (x-); например,

        ```
        x-mydocumentclass (4096=application/x-mydocumentclass)
        ```
   - Для каждого нового типа MIME повторите 2b и 2c, как требуется.

Совет: Контент-серверы OnDemand определяют поток типа MIME по расширениям файлов, а не по цифровым значениям классов содержимого.

Глава 3. Использование возможностей клиента администратора EIP

В этом разделе объясняются некоторые общие задачи, выполняемые администраторами EIP.

Создание объединенного поиска

Объединенный поиск - это запрос клиентской программы с поиском данных на одном или нескольких контент-серверах. EIP предлагает средства создания шаблонов для объединенного поиска. Так как каждый контент-сервер сохраняет и организует информацию по-своему, шаблон поиска должен учитывать эти различия для каждого сервера. Для поиска на контент-серверах шаблон поиска отображает объекты объединения и их атрибуты объединения на собственные объекты и атрибуты.

Создание объединенных поисков включает:

- Определение соединений с контент-серверами с использованием соединителей EIP.
- Создание объектов объединения
  - Определение объектов объединения
  - Создание атрибутов объединения
  - Отображение атрибутов объединения на собственные атрибуты
  - Назначение параметров
- Создание шаблонов поиска
  - Определение шаблона поиска
  - Определение критериев поиска
  - Определение параметров шаблона
  - Назначение доступа пользователям клиентов

EIP Версии 8.2 предлагает два мастера, который значительно облегчают создание объектов объединения и шаблонов поиска. Мастер по объектам объединения включает перечень серверов, который можно фильтровать для облегчения поиска собственных атрибутов. Он также устанавливает правильные параметры по умолчанию для атрибутов объединения, уменьшая возможность их неправильного конфигурирования. Мастер по шаблонам поиска помогает создавать шаблоны поиска. Он также поможет спроектировать внешний вид и действие критериев поиска и выводов результатов. Мастер даже позволяет предварительно просмотреть, как могут выглядеть шаблоны поиска в вашей клиентской программе. Кроме того, для тех, кто предпочитает объекты объединения и шаблоны поиска для EIP Версия 7.1, есть дополнительные диалоговые окна для создания таких объектов.

Все мастеры, диалоговые окна и поля описаны в электронной справке EIP.

Определение серверов

Чтобы подключиться к серверу и обновить перечень этого сервера, необходимо сначала определить сервер. Если вы щелкнете правой кнопкой мыши по значку сервера и выберете Новый, клиент выведет все соединители, поддерживаемые EIP. Для определения сервера необходимо знать некоторую основную информацию о соединителях:

- Какие соединители выбрал установщик? Установленные соединители перечислены в файле конфигурации `cmbcs.ini`. На сервере Windows путь по умолчанию *x*`:\Program Files\IBM\CMgmt`. В системах AIX или Solaris расположение файлов `cmbcs.ini` спросите у администратора системы.
- Выбрал установщик локальный или удаленный соединитель? Файл `cmbcs.ini` содержит локальные и удаленные типы соединителей.
- Если система сконфигурирована для RMI, запущен ли сервер RMI? Чтобы запустить RMI на локальном сервере RMI, выберите **Пуск → Программы → IBM Enterprise Information Portal for Multiplatforms 8.2 → Запустить серверы RMI**. Если ваша система использует удаленный реестр RMI, посмотрите в файле `cmbsvclient.ini`, на каком удаленном сервере установлены соединители RMI. Дополнительную информацию можно узнать у администратора сервера RMI.
- Если при установке EIP был установлен соединитель CM для AS/400, какая информация была включена в сетевую таблицу `frnolint.tbl`? Таблица AS/400 `frnolint.tbl` находится в `%CMBROOT%`.
- Если вы определяете удаленные контент-серверы, содержащие реляционные базы данных, например, Content Manager Версии 8 и DB2, DataJoiner и Information Catalog, надо внести в каталог или добавить базу данных с той рабочей станции, на которой используется клиент.

Ниже в списке описаны стандартные действия, которые необходимо выполнить при определении сервера:

1. Щелкните правой кнопкой мыши по папке Серверы и выберите **Новый**.
2. Выберите сервер из списка. Появится окно Новый сервер.
3. Введите имя сервера и описание в поле Имя сервера на вкладке Общие (сведения). Для некоторых серверов необходимо ввести только имя базы данных. Для других серверов необходимо ввести полное имя сервера, на котором установлена база данных.
4. Задайте, если требуется, параметры инициализации. Для некоторых серверов необходимо вводить параметры инициализации, такие как строка соединения и строки конфигурации. Для других серверов требуется указать лишь имя базы данных.
5. Нажмите кнопку Проверить сервер. EIP регистрируется на некоторых серверах, используя ID пользователя и пароль, введенные вами для запуска клиента администратора. Если для сервера необходим другой ID пользователя и пароль, EIP предложит ввести правильный ID пользователя и пароль специально для определяемого вами контент-сервера.

Совет: Можно также определить новый тип контент-сервера, однако тогда необходимо обеспечить классы соединителей Java или C++ и класс определения сервера для нового типа сервера. Понадобится также соединитель Java для составления перечня серверов. Инструкции по добавлению контент-серверов можно посмотреть в книге *Workstation Application Programming Guide* и в электронный справочник API.

Если конфигурация контент-сервера оказалась неправильной, смотрите в книге *Сообщения и коды* дополнительную информацию об устранении неисправностей или о действиях для полученного вами сообщения об ошибке.

Можно также обратиться за помощью к администратору сервера, к которому вы собираетесь подсоединиться.

Рекомендации по определению серверов

В этом разделе даются рекомендации, которые помогут выполнить начальное определение сервера.

Соединение с (реляционными) базами данных DB2

Этот раздел содержит информацию, касающуюся серверов DB2, DataJoiner, JDBC, ODBC, Information Catalog и Content Manager Версии 7 или 8.

- **Внимание:** Прежде чем определять сервер, необходимо внести в каталог каждую базу данных DB2. Для каталогизации баз данных можно использовать DB2 CCA или командную строку DB2. Более подробную информацию можно узнать у администратора DB2.
- В поле Имя сервера на вкладке Общие надо ввести имя базы данных, с которой вы хотите соединиться. Имя сервера надо вводить в верхнем регистре.
- При определении DB2, DataJoiner, JDBC, ODBC или Information Catalog щелкните по вкладке Параметры инициализации и введите Имя схемы, связанной с таблицами базы данных, с которыми вы соединяетесь, например `SCHEMA=ICMADMIN`.
- При определении сервера Content Manager Версий 7.1 или 8.2 нужно только ввести имя базы данных. Не меняйте на вкладке Параметры инициализации параметры по умолчанию.
- При определении сервера Content Manager Версии 7.1 на локальном диске в каталоге `x:\CMBROOT` должна находиться сетевая таблица `fronlint.tbl`. Сетевая таблица содержит информацию об имени хоста, номере порта и типе сервера, которая нужна EIP, чтобы найти библиотечный сервер Content Manager Версии 7.1 и зарегистрироваться на нем. При определении нескольких серверов Content Manager Версии 7.1 перед тем, как определить любой из серверов, необходимо сделать для него отдельную запись в файле `frnolint.tbl`.
- Чтобы соединиться с DB2 DataJoiner, убедитесь, что для Enterprise Information Portal для экземпляра базы данных, определенного в DB2 Universal Database, задан метод аутентификации сервер.
- Чтобы соединиться с DataJoiner 2.1, прежде чем определять сервер DataJoiner, надо загрузить программу связывания с сайта DataJoiner и связать базу данных DataJoiner.

Соединение с сервером текстового поиска

Чтобы определить сервер текстового поиска, надо сначала определить сервер Content Manager Версии 7.1, связанный с этим сервером текстового поиска.

Введите имя сервера текстового поиска в пункте ″Выберите имя связанного сервера Content Manager Версии 7.1″ из выпадающего списка. Этот список находится на вкладке Связанный сервер.

Чтобы EIP мог соединиться с сервером Content Manager Версии 7.1 и сервером текстового поиска, они должны работать.

Соединение с несколькими серверами Content Manager for AS/400

Если вы используете несколько серверов AS/400, надо определить дополнительные серверы в сетевой таблице. Эта сетевая таблица (`frnolint.tbl`) находится в *x*`:\<cmbroot>`. Для нового сервера введите его имя, тип соединения (например, TCP/IP), имя хоста, порт и тип сервера. Для первого сервера в процессе установки устанавливающий вводит значения сервера, имени хоста и порта; эти значения заносятся в таблицу `frnolint.tbl`.

Ниже приведен типовой пример информации в таблице `fronlint.tbl` :

```
/*  Сетевая таблица VI/400 */
SERVER: VI400 REMOTE TCPIP
        HOSTNAME = vi400
        PORT    = 29000
        SERVER_TYPE = FRNLS400
```

Конфигурирование соединителя Extended Search

Информация, вводимая для определения сервера Extended Search, зависит от двух факторов:

- От типа Web-сервера, на котором был установлен Extended Server - Web-сервер Domino, WebSphere, IIS.
- От номера порта, определенного для Web-сервера, на котором был установлен Extended Search.

При определении соединителя Extended Search:

1. В поле Имя сервера на вкладке Общие введите полное имя хоста Web-сервера, на котором был установлен Extended Search.
2. На вкладке Параметры инициализации введите в поле Номер порта значение 80, если во время установки Extended Search устанавливающий оставил для номера порта Web-сервера параметры по умолчанию.
3. В поле ID программы введите `Demo`. Введите имя, как показано.
4. В поле Пароль введите `Demo`.
5. В поле Дополнительные параметры:
   a. Не меняйте две точки с запятой, если знаете, что Extended Search был установлен на Web-сервере Domino и что во время установки были использованы параметры по умолчанию для номера порта Web-сервера и Extended Search.
   b. Ниже в разделе содержится информация о том, как изменить поле Дополнительные параметры для серверов Extended Search, если при установке были заданы пользовательские параметры.

Если соединитель Extended Search сконфигурирован с сервером прикладных программ WebSphere, если номер порта Extended Search НЕ 6001 или если имя сервера Extended Search НЕ совпадает с именем Web-сервера, нужно сконфигурировать соединитель, задав правильные относительные пути сервлета ES, номер порта ES и имя сервера ES.

Если Extended Search сконфигурирован с WebSphere, если номер порта Extended Search НЕ 6001 или если имя сервера Extended Search НЕ совпадает с именем Web-сервера, можно создать файл конфигурации, например `desclient.cfg`, выполнив следующие действия:

Перейдите в каталог, где находятся ваши программы или примеры. Создайте файл конфигурации, например `desclient.cfg`. Этот файл не поставляется с Enterprise Information Portal.

В файл `desclient.cfg` добавьте запись `DESHOSTNAME=(ES host name)`

Если ES сконфигурирован с сервером прикладных программ Domino, задайте `DESREQURI=/servlet/ESAdmin`. Если ES сконфигурирован с сервером прикладных программ WebSphere, задайте `DESREQURI=/lotuskms/ESAdmin`, где *servlet* - путь каталога на HTTP-сервере, поддерживающем соединитель Extended Search.

Если сервер прикладных программ - WebSphere, для `DESREQURI` надо задать не `/servlet/ESAdmin`, а `/lotuskms/ESAdmin`

Если вы собираетесь искать источники ES из минимальных или обычных клиентов, определите дополнительный параметр

```
"DESCFGPATH=<полный путь к desclient.cfg>"
```

в диалоговом окне определения сервера DES в клиенте администратора.

Если вы хотите запустить примеры ES, передайте полный путь `desclient.cfg` в аргументах командной строки.

Пример 1:

```
TConnectDES es.stl.ibm.com  user password
PORT=80;DESAPPID=Demo;DESAPPPW=password;DESCFGPATH
=<полный путь к desclient.cfg>;
```

Пример 2:

```
java TConnectDES es.stl.ibm.com  user password
PORT=80;DESAPPID=Demo;DESAPPPW=password;DESCFGPATH=<полный путь к
desclient.cfg>;
```

Определение сервера каталога данных

Прежде, чем вы определите сервер каталога данных, его надо внести в каталог. В поле Имя сервера введите имя сервера, например `SAMPLE1`. На вкладке Параметры инициализации введите `SCHEMA=<Имя схемы, связанной с SAMPLE1>`.

Определение сервера OnDemand

Чтобы можно было определить сервер OnDemand, должны работать демоны сервера OnDemand и библиотечного сервера. Перед тем, как определить сервер в EIP, можно послать запрос ping на сервер OnDemand, чтобы проверить, работают ли сервер и демон.

На вкладке Общие введите полное имя хоста для сервера, где был установлен OnDemand.

На вкладке Параметры инициализации введите номер порта, назначенного при установке сервера OnDemand. Если при установке OnDemand было выбрано значение порта по умолчанию 0, введите в поле Номер порта `0`. Если был выбран другой номер порта, введите знак # и этот номер порта. Например, на сервере Windows для OnDemand мог быть выбран порт `# 5000`.

Если вы определяете сервер OnDemand, который установлен на сервере AS/400 Версии 4, надо ввести в поле Дополнительные параметры следующие данные: `STATECONNECT=#1`.

Если вы определяете сервер OnDemand, который установлен на сервере OS/390 Версии 2.1, введите пользовательский номер порта, указанный при установке OnDemand на сервер OS/390 Версии 2.1.

OnDemand требует, чтобы гнездо было активировано во время соединения.

Определение сервера текстового поиска

Чтобы определить сервер текстового поиска, надо сначала определить сервер Content Manager Версии 7, связанный с этим сервером текстового поиска.

Введите имя сервера текстового поиска в поле Имя сервера на вкладке Общие. Выберите имя связанного сервера Content Manager Версии 7.1 из выпадающего списка на вкладке Связанный сервер.

Чтобы EIP мог соединиться с сервером Content Manager Версии 7.1 и сервером текстового поиска, они должны работать.

Определение сервера Domino.Doc

В поле Имя сервера введите путь к имени сервера и имени библиотеки сервера Domino.Doc. Например: `oakley/DominoDoc1/Lib.nsf`.

Если вы используете локальные соединители, надо установить на рабочей станции с клиентом EIP Domino Doc Desktop Enabler. При использовании RMI необходимо установить Domino Doc Desktop Enabler на сервере RMI. Версия Domino Doc Desktop Enabler должна совпадать с версией сервера Domino Doc.

Не изменяйте две точки с запятой на вкладке Параметры инициализации.

Определение ImagePlus для сервера OS/390

Чтобы иметь возможность подключиться к серверу ImagePlus for OS/390, при его определении надо установить указанные ниже параметры. Примеры значений приводятся ниже:

- FAF Port Number (Номер порта FAF): *3061*
- FAF Application ID (ID программы FAF): *01*
- FAF Protocol (Протокол FAF): *4000*
- FAF IP Address (IP-адрес FAF): *9.67.43.83*
- Object Distribution Manager CICS (Менеджер распределения объектов CICS): *4000*
- Object Distribution Manager IP Address (IP-адрес менеджера распределения объектов): *9.67.43.83*
- Object Distribution Manager Port Number (Номер порта менеджера распределения объектов): *3082*
- Object Distribution Manager Terminal ID (ID терминала менеджера распределения объектов): *оставьте это поле пустым*
- Additional parameters (Дополнительные параметры): *FAFSITE=CS61;;*

Использование трассировки в Content Manager ImagePlus for OS/390

Если не удается установить соединение с сервером Content Manager ImagePlus for OS/390, возможно, трассировка поможет решить проблемы. Если вы установили соединитель для Content Manager ImagePlus for OS/390, можно включить трассировку для ImagePlus for OS/390, отредактировав файл `eypapi.ini` в каталоге `cmbroot`.

Файл `eypapi.ini` содержит следующие строки:

```
; Path where the IPFAF files are stored
;    (MUST NOT have a trailing '\')
; -- default is the <ROOT Directory>\
;
IPFAFPath=d:\cmbroot
; Flag for Logging (EYPLmmdd.LOG files)
;        --  default is Logging OFF (0)
; --  0  All Logging OFF
; --  1  Log files created only error conditions logged
; --  2  Log files created all conditions logged
;
Logging = 0

   ;-------------------------------------------------------
;
; Flag for Logging the FAF Parameters Types created by APIs
;        --  default is Logging OFF (0)
```

```
;        --  0  Parameter types Not logged
;        --  1  Log Faf Parameter Types
;
FafTypeLogs = 0
```

IPFAFPath

Задает каталог, куда записываются журналы. Файлы журналов называются: EYP*mmdd*.LOG

где *ммдд* - месяц и день создания журнала.

Logging

Задает условия, при которых создается журнал.

0 Не создавать журнал. 0 - это значение по умолчанию.

1 Созданные файлы журнала содержат только записи об ошибках.

2 Созданные файлы журнала содержат все записи.

FafTypeLogs

Задает запись в журнал для типов параметров FAF, создаваемых API.

0 Не записывать типы параметров; значение по умолчанию - 0.

1 Записывать типы параметров FAF.

Работа с соединителем OnDemand: настройка TCP/IP и гнезда

Известная проблема в системе Windows может повлиять на производительность при соединении с сервером OnDemand. При повторении поиска на сервере OnDemand многие гнезда Windows открываются и закрываются. В Windows есть два параметра по умолчанию, которые могут повлиять на рост трафика между EIP и сервером OnDemand:

- Когда программа закрывает гнездо Windows, Windows переводит порт для гнезд в состояние TIME_WAIT на 240 секунд; на протяжении этого времени порт нельзя использовать снова.
- В Windows одна программа не может использовать более 5000 портов.

Чтобы избежать возможных проблем, измените значения срока ожидания и количества портов при помощи редактора реестра Windows.

- Измените время ожидания с 240 секунд на меньшее (допустимый диапазон значений - от 30 до 300 секунд). Это время задается ключом HKEY_Local_Machine\System\ CurrentControlSet\services\Tcpip\ Parameters\TcpTimedWaitDelay.
- Измените максимальное количество портов с 5000 (значение по умолчанию) на большее (допустимый диапазон значений - от 5000 до 65534). Это количество задается ключом HKEY_Local_Machine\System\CurrentControlSet\ services\Tcpip\Parameters\MaxUserPort

Дополнительную информацию о TcpTimedWaitDelay и MaxUserPort смотрите в документации Windows.

Работа с соединителем Extended Search

В этом разделе описываются изменения соединителя Extended Search в EIP Версии 8.2.

Заданная пользователем национальная версия поддерживается путем передачи значения национальной версии в ключе DESLOCALE. Если вы вызываете

соединитель ES непосредственно, можно передать это значение в виде пары в командной строке, Это значение можно задать на странице **Дополнительных** аргументов для свойств ES.

Замечание: программное обеспечение сервера Extended Search входит в поставку EIP Версии 8.2.

Создание объектов объединения

После определения соединений с контент-серверами следующее действие по созданию объединенного поиска - создание объекта объединения, который станет строительным блоком для шаблонов поиска. В этом разделе объясняется, что такое объект объединения и как использовать мастер по созданию объектов объединения.

Что такое объект объединения

В большинстве случаев пользователи клиентских программ не захотят искать информацию последовательно на каждом из серверов. Вместо этого они предпочтут запустить один единственный объединенный поиск. Шаблоны поиска позволяют пользователям клиентских программ объединить запрос к отдельным серверам в единый запрос. Как администратор EIP, вы можете создать подобные шаблоны поиска для использования в клиентских программах. Прежде чем создать шаблон поиска, необходимо сначала создать объекты объединения, отображающие объединенные атрибуты на собственные атрибуты на контент-серверах.

Например, DB2 хранит информацию в таблицах, в столбцах которой записаны атрибуты информации, хранящейся в таблице. Таблица под названием `Customer_Demographics` может содержать столбцы `Name`, `Pol_Number`, `Address`, `Phone` и `Occupation`.

С другой стороны, в Content Manager используются элементы, типы элементов и атрибуты, а не таблицы и столбцы. Та же информация, что и в DB2 может храниться в объекта под названием `CustInfo`. У него могут быть атрибуты - `CustName`, `Acct`, `HomeAddres`, `HomePhone` и `Job`. В этих случаях аналогичная информация хранится и идентифицируется по-разному.

EIP решает проблему учета всевозможных способов хранения аналогичной информации на контент-серверах. Объекты объединения отслеживают для вас такую информацию. Сам объект объединения в действительности не хранит данные; он хранит метаданные о том, каким образом каждый контент-сервер хранит данные. При создании объекта объединения вы отображаете все его атрибуты на соответствующие собственные атрибуты на тех контент-серверах, к которым будет обращаться запрос.

В приведенном выше случае можно создать объект объединения под названием `Policy_Info` с атрибутами объединения `Policy_Name`, `Policy_Number`, `Home_Address` и `Job_Title`. После этого можно отобразить атрибуты объединения на каждый из соответствующих собственных атрибутов.

EIP может сгенерировать перечень серверов, который содержит эту информацию. Эта информация содержится в перечне серверов, а мастер по созданию объектов объединения позволяет вам получать перечень серверов, который можно отфильтровать на контент-серверах. Фильтрация доступна только при использовании мастера по созданию объектов объединения. Если объект объединения создается

вручную (без использования мастера), фильтрацию нельзя использовать. Создав перечень серверов, можно начинать отображать объекты объединения на собственные атрибуты.

Однако недостаточно лишь отобразить объекты объединения на собственные атрибуты. Каждый собственный атрибут может также обладать различными свойствами. Атрибуты могут: (1) допускать пустые значения, (2) допускать поиск, (3) допускать изменения и (4) допускать текстовый поиск. В зависимости от типа данных, которые вы выберете, вам, возможно, надо будет указать опции длины, точности, масштаба, минимального и максимального значений данных.

При определении этих свойств их нельзя сделать более жесткими, чем свойства, уже определенные собственными атрибутами, отображенными на атрибуты объединения. Мастер устанавливает по умолчанию свойства в соответствии с этим критерием. Если после настройки свойств для атрибута объединения вы решите вернуться к параметрам по умолчанию, предлагавшимся мастером, вы все еще можете выбрать параметры по умолчанию.

Итак, атрибуты объединения отображаются на соответствующие собственные атрибуты на нескольких контент-серверах. Свойства каждого атрибута объединения' включают все свойства собственных атрибутов. После создания объекта объединения вы получаете путь к информации, хранящейся на разных контент-серверах. Объекты объединения можно также использовать для создания шаблонов поиска для определенных запросов.

Использование мастера по созданию объектов объединения

В EIP Версии 8.2 появился новый мастер по созданию объектов объединения. Для создания объекта объединения можно использовать те же диалоговые окна, что и в EIP Версии 7.1 или более ранних, но мастер значительно облегчает процесс создания объекта объединения.

Чтобы создать объект объединения при помощи мастера:

1. **Определите объект объединения** Определите имя объекта объединения и опишите объект объединения. Можно также определить, будет ли объект объединения допускать текстовый поиск.
2. **Определите атрибуты объединения** Определите имена атрибутов объединения и измените атрибуты объединения.
3. **Отобразите атрибуты объединения** Отобразите атрибуты объединения на собственные атрибуты. Предлагаемые средства позволяют получить перечень серверов, выбрать собственные атрибуты, которые надо отобразить, а позже - изменить отображения.
4. **Определите свойства** Определите свойства для каждого объекта объединения. Можно настроить свойства или принять параметры по умолчанию.
5. **Подтвердите объект объединения** Просмотрите еще раз параметры, выбранные для объекта объединения. Можно вернуться к предыдущим панелям и изменить параметры. Когда закончите, нажмите кнопку **Готово**.

Эти шаги соответствуют шагам, предлагаемым мастером. Дополнительную информацию по использованию мастера смотрите в электронной справке EIP.

Создание текстовых индексов объединения

Механизм текстового поиска может быть встроен в контент-сервер Content Manager Версии 7.1 или более ранней, в этом случае вы можете автоматически индексировать, искать и получать текстовую информацию, хранящуюся в Content Manager. Пользователи могут искать документы по словам или словосочетаниям. Сервер текстового поиска поддерживает как наборы однобайтных, так и наборы двухбайтных символов.

При использовании серверов Content Manager Версии 7.1. или более ранних с Механизм текстового поиска можно создавать текстовый индекс объединения. Затем вы отображаете текстовый индекс объединения на индексы текстового поиска Content Manager на серверах текстового поиска Content Manager.

Когда создаете индекс текстового поиска объединения, можно задать для него возможность комбинированного поиска, то есть поиска как по собственным текстовым индексам, так и по собственным атрибутам. Задавая возможность комбинированного поиска для текстового индекса объединения, вы также отображаете этот индекс на объект объединения. Затем вы отображаете собственные атрибуты и соответствующие атрибуты объединения, на собственные индексы текстового поиска на сервере текстового поиска.

Создание шаблонов поиска

После создания объекта объединения необходимо создать шаблон поиска. Помните, что шаблон поиска использует объект объединения как карту, где хранится информация. При создании шаблона поиска нужно определить, что вы собираетесь искать, что вы хотите сделать с результатами поиска и кто может пользоваться данным шаблоном. В одном шаблоне поиска можно использовать только один объект объединения, но один объект объединения можно использовать в нескольких шаблонах. Можно также задать в качестве критериев поиска любое сочетание атрибутов объекта объединения. Для создания шаблона поиска выполните с помощью мастера по шаблонам поиска следующие действия:

1. Определите шаблон поиска
2. Определите критерии поиска
3. Определите параметры поиска
4. Назначьте привилегии доступа

Эти шаги соответствуют шагам в мастере по шаблонам поиска. Подробную информацию по процессу создания шаблонов поиска смотрите в электронной справке EIP.

Определите шаблон поиска

После того, как вы запустите мастер, он предложит определить шаблон поиска. Будьте готовы:

- Дать имя и описание для шаблона поиска
- Выбрать объект объединения для шаблона поиска. **Ограничение**: В одном шаблоне поиска можно использовать только один объект объединения.
- Выберите текстовый индекс объединения, если он требуется

Совет: Переключатель текстового индекса объединения следует применять только при использовании механизма текстового поиска для Content Manager Версии 7.1 или более ранней. Если для поиска текста используется TIE DB2, этот поиск считается параметрическим и конфигурируется в шаблоне поиска соответствующим образом.

Определите критерии поиска

После того, как вы определите шаблон поиска, мастер предложит:

1. Выбрать тип поиска - атрибут или документ. Тип документ доступен, только если на предыдущем шаге был выбран текстовый индекс объединения.
2. Ввести имя для критерия поиска
3. Выбрать атрибут объединения
4. Выбрать операции из списка доступных
5. Задать строку поиска по умолчанию (только при поиске документов)

Мастер предлагает выпадающее меню со списком всех атрибутов объединения, связанных с выбранным объектом объединения. Эти атрибуты образуют критерии поиска для шаблона поиска. Мастер также предоставляет список возможных операций.

Совет: можно создать несколько критериев поиска в одном шаблоне или удалить существующие критерии из шаблона.

Определить параметры поиска

Эта панель позволяет определить параметры поиска, параметры критериев и параметры вывода по умолчанию. У каждого из этих параметров есть значение по умолчанию, которое можно изменить. Для изменения параметра нажмите соответствующую кнопку.

Окно Параметры по умолчанию позволяет:

- Задать, что делать, если при попытке пользователя применить шаблон сервер окажется недоступен.
- Определить символы подстановки для параметрического поиска
- Задать имя папки для сохранения результатов поиска
- Выбрать, как использовать критерии при поиске: все (AND) или любой (OR)

Окно Параметры критериев позволяет задавать порядок критериев поиска, порядок столбцов вывода результатов, заголовки столбцов и их ширину.

Окно Параметры вывода дает способ задать выводимые значения для результатов поиска. Например, если день недели на одном из серверов записан как Понедельник, а на другом - Пн, вы можете решить использовать при вывода результатов поиска значение Понедельник для обоих серверов.

Назначить привилегии

Надо задать не только где искать (объекты объединения), что искать (критерии поиска) и как выводить результаты (параметры вывода), но и кто может использовать шаблоны поиска.

Окно Назначение привилегий мастера по шаблонам поиска дает возможность задания доступа к шаблонам существующим пользователям или группам пользователей.

Назначение привилегий пользовательского доступа для шаблона поиска не дает этому пользователю доступа к контент-серверу, отображенному на шаблон. Пользователи должны соответствовать требованиям безопасности каждого контент-сервера в отдельности. Используя списки управления доступом и управление

пользователями, перед предоставлением пользователям доступа к шаблонам поиска убедитесь, что у пользователей есть необходимые привилегии.

При использовании мастера для поиска пользователей или групп пользователей EIP возвращает только пользователей, обладающих соответствующим доступом к запрошенным контент-серверам.

Глава 4. Как управлять доступом пользователей

Пользователь не может получить доступ к системе EIP, если у него нет ID пользователя, пароля или набора привилегий. Однако прежде чем создавать пользователей и предоставлять им привилегии, надо решить, у кого будет доступ к системе, и что им потребуется для работы. Было бы нежелательно, чтобы право удалять объект принадлежало пользователям, которые не понимают, к каким последствиям может привести удаление этого объекта. С другой стороны, не хотелось бы, чтобы недостаточность привилегий мешала пользователям выполнять их работу. Поэтому прежде чем назначать пользователям привилегии, надо определить типы задач для каждой должности.

Когда пользователи создают объект в системе EIP, они должны задать, какой доступ к нему получат другие пользователи. Пользователи, создающие объект, должны определять, кто сможет к нему обращаться и какие операции с ним проводить. Такое определение в системе EIP называется списком управления доступом - ACL (Access Control List).

Создание ID пользователей и паролей

Если вы хотите, чтобы ID пользователя, определенный в клиенте администратора системы, использовался и для аутентификации DB2, этот ID пользователя должен удовлетворять правилам именования DB2. Правила именования DB2 применяются для ID пользователей, которые будут использоваться как ID администраторов или как ID пользователей для соединения. Нельзя использовать следующие слова:

- USERS
- ADMINS
- GUESTS
- PUBLIC
- LOCAL
- Все зарезервированные слова SQL, перечисленные в Справочнике по SQL.

ID пользователя не может начинаться со следующих символов:

- SQL
- SYS
- IBM

Можно использовать следующие символы:

- Буквы от A до Z. **Ограничение:** В некоторых операционных системах допускаются регистрочувствительные ID пользователей и пароли. Чтобы узнать, так ли это в вашей операционной системе, посмотрите документацию на операционную систему.
- Цифры от 0 до 9
- #
- $

Ограничение: ID пользователя не может содержать более 30 символов.

О полномочиях администратора DB2

При регистрации в клиенте администратора системы вы проходите два уровня аутентификации: один - на уровне базы данных и второй - на уровне продукта. При включенной возможности управления доменами есть два вида администраторов: старший администратор и подадминистратор. В общем случае только старшие администраторы имеют доступ к клиенту администратора системы.

Старшие администраторы должны обладать привилегиями DB2: требуются привилегии db2admin, то есть полные привилегии администратора DB2. Этот ID пользователя должен быть определен в операционной системе с привилегией db2admin. Пароль этого ID пользователя операционной системы используется для соединения с DB2 и для регистрации на библиотечном сервере. Пароль, определенный на библиотечном сервере, не используется. Привилегии Content Manager: Этот ID пользователя определяется на библиотечном сервере с полными привилегиями администратора Content Manager (″AllPrivs″), чтобы он мог выполнять все действия администратора.

Подадминистраторам не требуются привилегии DB2. Подадминистраторы управляют только определенными разделами на библиотечном сервере, поэтому регистрируются в клиенте администратора системы одним из двух способов:

- Если этот ID пользователя - ID пользователя операционной системы, его пароль для операционной системы используется для соединения с DB2 и для регистрации на библиотечном сервере.
- Если этот ID пользователя не является ID пользователя операционной системы, для соединения с DB2 используется шифрованная пара ID пользователя - пароль из файла cmbfedenv.ini (для Enterprise Information Portal) или cmbicmenv.ini (для Content Manager), а для регистрации на библиотечном сервере используются ID пользователя и пароль, введенные в окне регистрации.

Дополнительную информацию о регистрации на библиотечном сервере смотрите в следующем разделе.

Подадминистраторам требуются также привилегии EIP. Им требуется привилегия администратора домена для выполнения всех административных действий в домене.

Соединение с DB2 при помощи INI-файлов

Каждая запись в INI-файле содержит имя библиотечного сервера и шифрованные ID пользователя и пароль для соединения с DB2. Эти шифрованные ID пользователя (называемый ID пользователя для соединения) и пароль определяются во время установки продукта. ID пользователя для соединения не должен совпадать с ID пользователей администраторов системы. Для соединения с DB2 Enterprise Information Portal использует файл cmbfedenv.ini, а Content Manager - cmbicmenv.ini. По умолчанию ID пользователя для соединения - ICMCONCT. Во время установки пароли для библиотечного сервера и менеджера ресурсов находятся в трех местах: Файл cmbicmenv.ini содержит ID пользователя и пароль для доступа к библиотечному серверу. Доступ к базе данных, где находятся библиотечный сервер и менеджер ресурсов, определяется операционной системой. Файл ICMRM.properties содержит ID пользователя и пароль для менеджера ресурсов.

Если используется INI-файл, то есть ID пользователя не является ID пользователя операционной системы, этот ID пользователя и ID пользователя для соединения, определенные в этом INI-файле, должны существовать на библиотечном сервере.

ID пользователя для соединения должен быть определен на библиотечном сервере и в операционной системе. Он должен обладать привилегией UserDB2Connect. Чтобы изменить ID пользователя для соединения и пароль в INI-файле, выберите **Инструменты --> Изменить ID/пароль для базы данных** в окне клиента администратора.

Изменение пароля библиотечного сервера и администратора системы для менеджера ресурсов

Если нужно изменить пароль для менеджера ресурсов, необходимо изменить изменить пароль для регистрации библиотечного сервера в менеджере ресурсов и пароль администратора системы для менеджера ресурсов. **Внимание:** Изменяя пароли для регистрации библиотечного сервера и администратора системы в менеджере ресурсов, выполните следующие действия в указанном порядке:

1. Зарегистрируйтесь в клиенте администратора системы.
2. Раскройте дерево Менеджер ресурсов.
3. Выберите менеджер ресурсов, который хотите изменить, и раскройте его дерево.
4. Выберите Определения серверов и затем Свойства. Откроется окно Панель сервера.
5. Измените пароль в поле Пароль.
6. Нажмите кнопку OK.
7. Щелкните правой кнопкой мыши по менеджеру ресурсов, который вы раскрыли на шаге 2, и выберите Свойства. Откроется окно Свойства менеджера ресурсов.
8. Измените пароль в поле Пароль и нажмите кнопку OK.

Изменение паролей доступа к базе данных

Если нужно изменить пароли доступа к базе данных, необходимо изменить изменить пароль операционной системы для соединения с базой данных и файл ICMRM.properties, чтобы менеджер ресурсов мог узнать новый пароль.

Чтобы изменить пароль операционной системы для соединения с базой данных:

1. Запустите утилиту Пользователи и пароли (в соответствии с вашей операционной системой).
2. Выберите ICMRM.
3. Выберите Задать пароль.
4. Введите новый пароль.

Чтобы изменить файл ICMRM.properties:

1. Откройте файл `ICMRM.properties`. По умолчанию его полный путь - *X*`:\WebSphere\AppServer\installedApps\icmrm.ear\icmrm.war\WEB-INF\classes\com\ibm\mm\icmrm\ICMRM.properties`, где *X* - буква диска, где установлен Content Manager.
2. Измените DBPassword, чтобы он совпадал с паролем операционной системы.
3. Сохраните файл `ICMRM.properties`.

Изменив пароль базы данных, нужно перезапустить базу данных или позволить ей выдать два или три сообщения об ошибках, после чего она сама воспримет новое значение.

Подробные инструкции по изменению паролей и других полей для менеджера ресурсов в клиенте администратора системы смотрите в справке клиента администратора системы.

Импорт пользователей из LDAP

LDAP поддерживает управление ID пользователя и паролем на уровне предприятия, а не на уровне отдельных систем. EIP использует три технологии LDAP: IBM Directory (старое название IBM SecureWay Directory), Windows 2000 Active Directory и Lotus Domino Directory Notes Address Book (NAB). Пароли пользователей находятся на сервере LDAP. При регистрации пользователя в или Enterprise Information Portal выполняется аутентификация ID пользователя и пароля и проверяются привилегии данного ID пользователя по профилю пользователя в базе данных EIP. LDAP может быть включен во время установки EIP. Если LDAP не был включен во время установки, его можно активировать в любое время.

Чтобы включить LDAP, выберите **Пуск → Программы → EIP for Mulitplatforms → Планировщик импорта ID пользователей LDAP** и затем запустите клиент администратора системы. Откройте окно Конфигурация LDAP (Инструменты --> Конфигурация LDAP). Включите переключатель Разрешить импорт пользователей LDAP и аутентификацию и задайте информацию сервера LDAP на странице Сервер.

Когда LDAP включен, можно импортировать пользователей, нажав кнопку LDAP в окне Новый пользователь. Это позволяет выборочно импортировать пользователей с сервера LDAP в EIP. Другой способ - импортировать пользователей в группы при помощи утилиты Планировщик импорта ID пользователей LDAP. Во время регистрации библиотечный сервер автоматически соединяется с сервером LDAP для аутентификации пользователя. Если сервер LDAP по какой-либо причине не может подтвердить правильность пароля пользователя, аутентификация будет неудачной.

Конфигурацию сервера LDAP можно изменить, перейдя в главное окно клиента администратора системы и выбрав **Инструменты -> Конфигурация LDAP**. Изменить текущий сервер LDAP можно также, запустив утилиту импорта реестра пользователей LDAP: из меню Пуск для EIP. Информацию о планировании для LDAP смотрите в книге *Планирование и установка Content Management*. Информацию о конфигурировании информации сервера LDAP в окнах клиента администратора системы смотрите в электронной справке клиента администратора системы.

Информацию о планировании для LDAP смотрите в книге *Планирование и установка вашей системы Content Management*. Информацию о реализации LDAP смотрите в электронной справке клиент администратора системы.

О привилегиях

Клиент администратора содержит группы привилегий, наборы привилегий и отдельные привилегии. При управлении совместной системой Content Manager/EIP привилегии будут общими для обеих частей клиента. Встроенные в этот клиент привилегии помогут вам упростить

Группа привилегий
Группа привилегий - это собрание задач пользователя, помогающее администраторам создавать новые наборы привилегий или использовать роли в диалоговом окне Набор привилегий.

Наборы привилегий
Наборы привилегий - это собрания ролей пользователей.

Привилегия
Привилегия представляет некоторое действие пользователя. Например:

Пример 1 - привилегии: Вы хотите предоставить привилегии ClientScan и ClientImport группе пользователей, которые обычно используют клиент только для сканирования и импорта документов в Content Manager. Если есть несколько пользователей, которые будут выполнять эту задачу, для них можно создать один ID пользователя (например, user1). Затем нужно связать привилегии ClientScan и ClientImport с ID пользователя User1. Затем вы помещаете пользователя User1 в группу Group1. Когда какой-либо конечный пользователь введет user1 для регистрации в своем клиенте и обращения к Content Manager, он сможет только сканировать и импортировать документы.

Пример 2 - группы привилегий: У вас есть группа опытных пользователей, которым требуются привилегии для выполнения всех типичных задач клиента. Вы можете создать ID пользователя (например, user2). Затем поместить пользователя user2 в группу (например, group2). Затем нужно назначить пользователю user1 группу привилегий ClientTaskAll. Когда какой-либо конечный пользователь введет user2 для регистрации в своем клиенте и обращения к Content Manager, он сможет выполнять все задачи, входящие в группу привилегий ClientTaskAll.

Пример 3 - наборы привилегий: У вас есть группа пользователей, которым требуется доступ только для чтения. Вы можете создать ID пользователя (например, user3). Затем поместить пользователя user3 в группу (например, group3). Затем нужно назначить пользователю user3 набор привилегий ClientUserReadOnly. Когда какой-либо конечный пользователь введет user3 для регистрации в своем клиенте и обращения к Content Manager, он сможет выполнять только задачи, входящие в набор привилегий ClientUserReadOnly.

Создание наборов привилегий

Когда вы планируете конфигурацию системы EIP, надо, помимо прочего, решить, какие пользователи будут иметь доступ к системе и какие полномочия они получат для объектов системы. В системе EIP доступ определяется с помощью привилегий.

Привилегия дает право доступа определенного рода к определенному объекту. Привилегии включают в себя такие права, как создание, удаление и выбор объектов, хранящихся в системе. Назначенная пользователю группа привилегий называется набором привилегий.

Первая задача управления доступом - создание наборов привилегий для пользователей. *Набор привилегий* определяет задачи или действия, которые могут выполнять пользователи. Наборы привилегий объединяют привилегии; они предназначены для определенных типов пользователей. Например, вы, возможно, захотите иметь один набор администраторов, управляющих сервером маршрутизации документов, а другой набор администраторов - для управления доменом. При регистрации администратора EIP проверяет его набор привилегий.

В клиенте администратора системы есть ряд заранее определенных привилегий, из которых можно составлять наборы привилегий. Создаваемые наборы привилегий можно затем предоставлять отдельным пользователям. Нельзя предоставить набор привилегий группе пользователей.

Как создавать группы привилегий

Группы привилегий аналогичны группам пользователей. Группу привилегий создают, чтобы собрать сходные привилегии, чтобы потом удобнее было найти нужные привилегии и включить их в некоторый набор привилегий. Например, если у вас есть две привилегии, которые вы предоставляете почти всем пользователям системы,

можно, вместо того чтобы проводить поиск в длинном списке привилегий при каждом создании набора привилегий, просто собрать эти две базовые привилегии в группу привилегий с именем BasicPrivs.

Назначение набора привилегий пользователю

В клиенте администратора системы есть ряд заранее определенных привилегий, из которых можно составлять наборы привилегий. Создаваемые наборы привилегий можно затем предоставлять отдельным пользователям. Нельзя предоставить набор привилегий группе пользователей.

Можно создавать имена привилегий, но нельзя создавать сами привилегии. Чтобы создать новые привилегии, которых нет в клиенте администратора системы, потребуется совместная работа с системным программистом.

Можно использовать наборы привилегий, поставляемые с EIP, или создавать собственные наборы.

Назначение ID пользователя набора привилегий с правом предоставления

Чтобы пользователи не могли создавать ID пользователей с большими привилегиями, чем у них самих, в EIP реализовано использование набора привилегий с правом предоставления. Когда вы задаете для ID пользователя набор привилегий с правом предоставления, вы даете ему полномочия создавать ID пользователей в рамках предоставленных ему привилегий. Например, ID пользователя можно дать некоторые привилегии системного администратора - привилегии, позволяющих управлять доменом. Возможно, однако, вы захотите, чтобы этот ID пользователя не имел привилегии создавать пользователей. При создании такого пользователя в поле набор привилегий с правом предоставления выберите ″Noprivs″. В результате этот ID пользователя сможет управлять доменом, но не сможет создавать в нем пользователей.

Назначение менеджера ресурсов для пользователей

Чтобы позволить пользователям обращаться к определенному менеджеру ресурсов, вы назначаете для этого менеджера ресурсов домен, к которому у пользователей есть доступ. Дополнительную информацию о задании доменов для менеджеров ресурсов смотрите в разделе “Назначение домена для менеджера ресурсов” на стр. 37.

Назначение собраний для пользователей

Чтобы позволить пользователям обращаться к собраниям, вы назначаете для собрания домен, к которому у пользователей есть доступ. Дополнительную информацию о назначении доменов для собраний смотрите в разделе “Назначение домена для собрания” на стр. 38.

Как создавать группы пользователей

Нередко пользователи выполняют одинаковую работу и одни и те же или близкие задачи; поэтому им нужны одни и те же права доступа к объектам системы. Можно сгруппировать пользователей, которым требуются одни и те же права доступа, создав группу пользователей. Нельзя включать одни группы пользователей в другие.

Группа пользователей - это просто удобный способ группировать отдельных пользователей с похожими задачами. Группе пользователей не назначается набор привилегий. Каждый пользователь в группе пользователей обладает своим

собственным набором привилегий. Группы пользователей упрощают создание списков управления доступом к объектам вашей системы.

Если вы включили домены, прежде чем включать ID пользователя в группу, проверьте, принадлежит ли эта группа пользователей определенному домену или общедоступному домену PUBLIC (дополнительную информацию о доменах смотрите в разделе “Управление доменами” на стр. 36). Убедитесь, что группа пользователей находится в домене, в который вы хотите поместить ID пользователя. Если вы хотите создать ID пользователя специально для домена, можете нажать кнопку **Новый пользователь** в окне Группа пользователей. После этого можно добавить создаваемого пользователя к группе пользователей, будучи уверенным, что этот пользователь попадет в тот же домен.

Как создавать списки управления доступом

Дайте пользователям привилегии, необходимые для выполнения их задач. Объекты, каждый в отдельности, имеют свои возможности управления доступом.

Список управления доступом (ACL) включает в себя один или несколько ID отдельных пользователей или групп пользователей и присвоенные им привилегии. Списки управления доступом (ACL) используются для управления доступом пользователей к объектам в системе EIP. Со списками управления доступом могут быть связаны такие объекты, как данные, записанные пользователями, типы и поднаборы типов элементов, рабочие списки и процессы.

Наборы привилегий определяют для отдельного пользователя максимальные возможности при использовании системы, а ACL ограничивает доступ отдельного пользователя к объекту. Наличие в ACL привилегии, не определенной в наборе привилегий пользователя, не предоставляет пользователю эту привилегию. Только пользователи, обладающие этой привилегией, могут пользоваться этой привилегией для объекта. Списки управления доступом только ограничивают доступ пользователей, они не дают дополнительных возможностей доступа. При работе с системой списки управления доступом обеспечивают дополнительный уровень защиты.

Назначение набора привилегий для списка управления доступом

С каждым ID пользователя, который вы добавляете в список управления доступом (ACL), надо связать набор привилегий. ID пользователя и набор привилегий определяют, какие пользователи получат доступ к объекту и какой это будет доступ.

Пользователи не получают доступа к объекту, если их нет в ACL. Чтобы добавить в ACL пользователя или группу пользователей, надо выбрать ID пользователя и набор привилегий для ACL и нажать кнопку **Добавить**. Для каждого заданного ACL вы найдете ID пользователей и группы пользователей в окне Список управления доступом. Вы можете изменять эту таблицу, добавляя и удаляя ID пользователей и группы пользователей. Дополнительную информацию о создании и изменении ACL смотрите в электронной справке клиента администратора системы.

Создание доменов

Домен - это раздел управляющей базы данных, которым управляет один или несколько администраторов. В домены входят ID пользователей, группы пользователей, списки управления доступом, наборы привилегий, менеджеры ресурсов и собрания SMS. Домены невидимы для пользователей, так что имя,

которое вы даете домену, должно быть осмысленным только для вас и системных администраторов, управляющих эти доменом. Пользователи не знают, что их доступ ограничен частью управляющей базы данных, то есть им известны только объекты внутри домена.

Домены ограничивают доступ пользователей и администраторов подразделом управляющей базы данных. Администратор с полными привилегиями для управляющей базы данных может передать часть своих привилегий управления другому администратору. Администратор с полными привилегиями - старший администратор - имеет доступ ко всем разделам управляющей базы данных, а администратор с ограниченными привилегиями - подадминистратор - только к разделу управляющей базы данных.

Домены ограничивают доступ подадминистратора к спискам управления доступом (ACL). Только старшие администраторы создают ACL, которые подадминистраторы могут использовать для добавления и удаления ID пользователей и групп пользователей. Подадминистраторы не могут создавать, изменять и удалять ACL.

Подадминистратор может выполнять некоторые обязанности старшего администратора, но только в своем домене. Создавая домены и задавая администраторов для управления ими, старшие администраторы могут переложить на подадминистраторов управление пользователями и задачами, уникальными для их домена, и сосредоточиться на системе в целом.

Прежде чем включить домены, примите во внимание, что:

- Включение доменов нельзя отменить
- Менеджеры ресурсов, собрания, ID пользователей и группы пользователей могут существовать в каждый момент времени только в одном домене
- Наборы привилегий и списки управления доступом могут существовать в каждый момент времени в нескольких доменах.
- За исключением общедоступного домена PUBLIC, домены не перекрываются
- Никакой объект, созданный в домене старшего администратора, нельзя удалить - ни сгенерированный системой, ни созданный пользователем.

Чтобы разрешить домены, войдите в меню Файл, выберите **Инструменты → Домены администраторов** и выберите **Включить домены администраторов**. Чтобы разрешение доменов вступило в силу, нужно перезапустить клиент администратора системы. Подробные инструкции о том, как настроить управляющую базу данных для доменов, смотрите в электронной справке клиента администратора системы.

Управление доменами

В зависимости от вашего набора привилегий вы управляете либо всей управляющей базой данных, либо ее определенным доменом. Администратор с полным доступом к управляющей базе данных называется старшим администратором. Подадминистратор имеет полный доступ к определенному домену.

Администраторы любого типа имеют возможность в своих доменах создавать, получать, изменять и удалять объекты, в том числе пользователей и собрания. Подадминистраторы видят и могут получать объекты в их доменах, а также получать списки объектов или объекты в общедоступном домене PUBLIC.

Доступ к доменам

Подадминистраторы не могут менять домен объекта. Однако они могут обращаться к содержимому своего домена, а также получать списки объектов и сами объекты из домена PUBLIC - домена совместного использования.

Старшие администраторы имеют доступ ко всем доменам в управляющей базе данных. Они могут создать объект и задать для него домен. Некоторые объекты, такие как наборы привилегий и списки ACL, создаются исключительно старшими администраторами, а подадминистраторы только используют их.

Подадминистратор может создавать, получать, изменять и удалять (CRUD) любые объекты только в своем домене.

Назначение домена для пользователя

Создавая ID пользователя, вы можете включить его в домен или оставить его в домене по умолчанию. Можно изменить домен, заданный для ID пользователя, после его создания - при помощи свойств пользователя.

ID пользователя в каждый момент может иметь доступ только к одному домену. Нельзя добавить пользователя в домен совместного использования PUBLIC.

Только у старших администраторов есть полномочия создавать домены и включать в них пользователей. У домена может быть несколько подадминистраторов, но только старший администратор может определять, кто будет этими администраторами, предоставляя им привилегии системных администраторов в пределах набора привилегий. Поле **Предоставить набор привилегий** в окне Новый пользователь или Свойства пользователя показывает, какие привилегии управления у подадминистратора есть в домене.

Назначение домена для группы пользователей

Включение группы пользователей в домен изменяет домен, заданный для каждого ID пользователя в этой группе. ID пользователя в каждый момент может иметь доступ только к одному домену. Поэтому всех ID пользователей, входящие в группу, которую вы добавите в домен, также будут включены в новый домен.

Группа пользователей в каждый момент может находиться только в одном домене. Можно добавить группу пользователей в домен совместного использования PUBLIC.

Назначение домена для набора привилегий

Все ID пользователей, которые вы добавляете в домен, должны иметь также связанные с ними привилегии. Если вы не включите связанные с ними наборы привилегий, пользователи не смогут выполнять свои задачи. Чтобы сделать наборы привилегий доступными для любых пользователей, лучше всего хранить их в домене PUBLIC - домене совместного использования.

Назначение домена для менеджера ресурсов

Вы можете ограничить доступ пользователя к определенным менеджерам ресурсов, включив их в определенный домен. Когда вы определяете новый менеджер ресурсов для управляющей базы данных, вы можете, если нужно, выбрать домен.

По умолчанию все менеджеры ресурсов назначаются в домен PUBLIC. Если же вы не хотите, чтобы все имели доступ к этому менеджеру ресурсов, надо включить его в некоторый домен. Если вы не находите подходящего домена, в который можно

поместить данный менеджер ресурсов, можно определить менеджер ресурсов, а потом создать нужный домен. Определив этот домен, откройте свойства менеджера ресурсов и выберите этот домен.

Назначение домена для собрания

Вы можете ограничить доступ пользователя к определенному собранию в менеджере ресурсов, включив собрание в определенный домен. Если менеджер ресурсов находится в домене PUBLIC, собрание можно включить в любой другой определенный домен. Но если менеджер ресурсов уже помещен в некоторый домен, нельзя включить собрание в другой домен, даже в домен PUBLIC.

Пользователю требуется доступ к менеджеру ресурсов, чтобы получить доступ к его собраниям, поэтому нельзя ограничивать доступ к менеджеру ресурсов, не налагая такие же ограничения на его собрания.

Перемещение пользователей из одного домена в другой

Может возникнуть ситуация, когда некоторого пользователя нужно будет переместить из одного домена в другой. При помощи поля **Описание** в окне Определение пользователя можно запомнить, к каким группам пользователей принадлежат конкретные пользователи. Возможно, это слегка упростит задачу.

Внимание: Данная задача отнимает много времени и при ошибке выполнения может привести к проблемам при обращении к системе. Чтобы изменить домен пользователя, надо быть старшим администратором.

Выполните следующие шаги, соблюдая особую внимательность:

1. Найдите все группы, к которым принадлежит пользователь.
2. Переместите все группы, к которым принадлежит пользователь, в домен PUBLIC, либо удалите из них данного пользователя.
3. Переместите в домен PUBLIC все менеджеры ресурсов, относящиеся к данному пользователю, а затем и все относящиеся к ним собрания.
4. Создайте (*не перемещая*) все наборы привилегий, связанные с данным пользователем, если их еще нет в домене назначения.
5. Создайте (*не перемещая*) все связанные с пользователем списки доступа в домене назначения, если их еще там нет.
6. Переместите пользователя в домен назначения: откройте свойства пользователя и измените его домен.
7. **Необязательно:** Можно переместить из домена PUBLIC в домен назначения группы и менеджер ресурсов, которые вы переместили на шаге 1, 2 и 3, но это можно делать, только если в исходном домене не осталось больше пользователей, связанных с перемещаемыми группами и менеджером ресурсов. В противном случае эти группы и менеджер ресурсов нужно оставить в домене PUBLIC, чтобы не нарушить их совместное использование для пользователей в других доменах.

Помните: Пользователь никогда не может находиться в домене PUBLIC. Пользователей нельзя использовать совместно.

Перемещение группы пользователей из одного домена в другой

Внимание: Эта задача отнимает много времени и при ошибке выполнения может привести к проблемам при обращении к системе. Чтобы изменить домен группы пользователей, надо быть старшим администратором.

Чтобы переместить группу пользователей в другой домен, выполните следующие действия:

- Если группа пользователей пустая, удалите эту группу из текущего домена, а затем воссоздайте ее и назначьте в домен назначения.
- Если группа пользователей не пустая, выполните следующие действия:
  1. Найдите всех пользователей, принадлежащих этой группе.
  2. Удалите группу из текущего домена, что приведет к удалению всех пользователей.
  3. Воссоздайте группу и назначьте ее в домен назначения.
  4. Добавьте во вновь созданную группу всех ее пользователей.

Перемещение менеджера ресурсов из одного домена в другой

Чтобы изменить домен менеджера ресурсов, надо быть старшим администратором.

Чтобы переместить менеджер ресурсов в другой домен, выполните следующие действия:

- Если менеджер ресурсов не содержит собраний, переместите его в домен назначения, открыв свойства этого менеджера ресурсов и изменив текущий домен на домен назначения.
- Если менеджер ресурсов содержит собрания, выполните следующие действия:
  1. Переместите менеджер ресурсов в домен PUBLIC.
  2. Переместите его собрания в домен назначения: откройте Свойства и выберите нужный домен назначения.
  3. Переместите в домен назначения менеджер ресурсов: откройте Свойства и выберите нужный домен назначения.

Перемещение собрания из одного домена в другой

Чтобы изменить домен собрания, надо быть старшим администратором.

Чтобы переместить собрание из одного домена в другой, выполните следующие действия:

1. Найдите менеджер ресурсов, которому принадлежит это собрание.
2. Переместите связанный с собранием менеджер ресурсов в домен PUBLIC.
3. Переместите собрание в домен назначения: откройте Свойства и выберите нужный домен назначения.
4. Переместите в домен назначения менеджер ресурсов: откройте Свойства и выберите нужный домен назначения.

Перемещение привилегии из одного домена в другой

Поскольку набор привилегий может располагаться в нескольких доменах, его можно добавить в домен назначения, не перемещая.

Перемещение списка управления доступом из одного домена в другой

Поскольку список управления доступом может располагаться в нескольких доменах, его можно добавить в домен назначения, не перемещая.

Глава 5. Управление исследованием информации

В этом разделе сначала объясняется, что такое исследование информации и как можно его использовать в деловой среде. Затем идут разделы, описывающие первые шаги исследования информации, принципы исследования информации, программу Information Structuring Tool, а в конце излагаются замечания по настройке производительности.

Что такое исследование информации?

Исследование информации Enterprise Information Portal - это ключевая технология, автоматизирующая многие операции извлечения и анализа информации и позволяющая компаниям с малыми затратами предоставить своим пользователям легкий доступ к нужной информации.

Первая цель исследования информации - получение из неструктурированного текста информации, с которой сможет работать компьютер. Современная технология еще не позволяет проводить полный анализ фактов, изложенных на естественном языке. Однако инструменты, в которых применяются технологии распознавания образов и эвристические алгоритмы, могут извлекать значимую информацию из текста произвольной структуры. Из документа извлекаются найденные в нем так называемые ″важные слова″, такие как имена, названия учреждений и географические названия, из которых составляется сводка для этого документа.

Исследование информации не только извлекает элементы информации из отдельного документа. Эта технология применяется при работе с обширными собраниями документов. Метафора ″исследование″ обозначает как процесс обнаружения информации, то есть поиска и выделения краткой информации из отдельных документов и сохранения этой информации в виде метаданных, так и процесс анализа распределений выделенных элементов по собранию документов, позволяющий обнаружить интересные явления, свойства и тенденции.

Службы исследования информации Enterprise Information Portal

Службы исследования информации Enterprise Information Portal предоставляют инфраструктуру для создания и сохранения информации, относящейся к отдельным документам и собраниям документов. Эта информация называется метаданными. Примеры информации, характеризующей содержимое документа и сохраняемой в виде метаданных:

- Заголовок
- Краткое содержание или сводка
- Имена, термины или выражения
- Категории, к которым относится документ

Отличие исследования информации от традиционного склада метаданных, связывающего документы с метаданными, в том, что исследование информации позволяет автоматически создавать метаданные, даже когда они не доступны в явном виде. Алгоритмы исследования и получения информации могут получать нужную информацию из огромных собраний документов - используя метаданные в процессе получения информации или же применяя к метаданным статистические модели,

чтобы найти интересные связи между документами, которые могут быть не очевидны при просмотре отдельных документов собрания.

Поскольку операции исследования и получения информации вместо содержимого оригинальных документов работают с четко определенным набором метаданных, скорость этих процессов можно значительно повысить, сохраняя метаданные в специальном хранилище - так называемом *складе данных исследования информации*. Это значительно повышает скорость доступа к метаданным, так как программы могут получать метаданные из одного хранилища и им не нужно снова обращаться к самим удаленным контент-серверам. Метаданные играют ключевую роль в поиске и получении информации и часто используются для уточнения результатов поиска, поэтому многие шаги могут выполняться вообще без обращения к контент-серверам.

Другое преимущество использования склада метаданных - данные, связанные с неким документом, хранятся в нем отдельно от самого документа. Очевидно, что для документов, для которых возможен доступ только для чтения (например, внешних документов в Web), вообще нельзя сохранять метаданные в том же месте, что и содержимое.

Службы исследования информации Enterprise Information Portal предоставляют следующие механизмы автоматического создания метаданных:

- **Категоризация** - относит документ к одной или нескольким категориям на основе определенной пользователем таксономии. Компонент категоризации содержит программу с графическим пользовательским интерфейсом, используемую для создания и поддержания таксономий; эта программа называется *Information Structuring Tool*.
- **Составление сводок** - выделяет наиболее важные фразы документа, на основе которых пользователь может решить, нужно ли читать весь документ. Пользователь может задать нужную длину сводки, влияя таким образом на баланс между сложностью выделенных метаданных и объемом информации в документе.
- **Распознавание языка** - определяет язык, на котором написан документ. Это полезный шаг предварительной обработки, применяемой перед другими службами исследования информации.
- **Извлечение информации** - автоматически распознает важные словарные элементы (такие как имена, термины и выражения) в тексте документов.
- **Кластеризация** - делит набор документов на группы похожих документов (кластеры). Кластеры автоматически выделяются из собрания документов.

Компоненты служб исследования информации

Программа исследования информации обычно включает в себя выполнение следующих задач:

1. Организация данных для просмотра и перемещения по ним
2. Обращение к разнородным источникам данных
3. Использовании операции расширенного поиска, выделяющей нужные данные для генерации прогноза или определения тенденций

Эти задачи исследования информации показаны на рис. 1 на стр. 43.

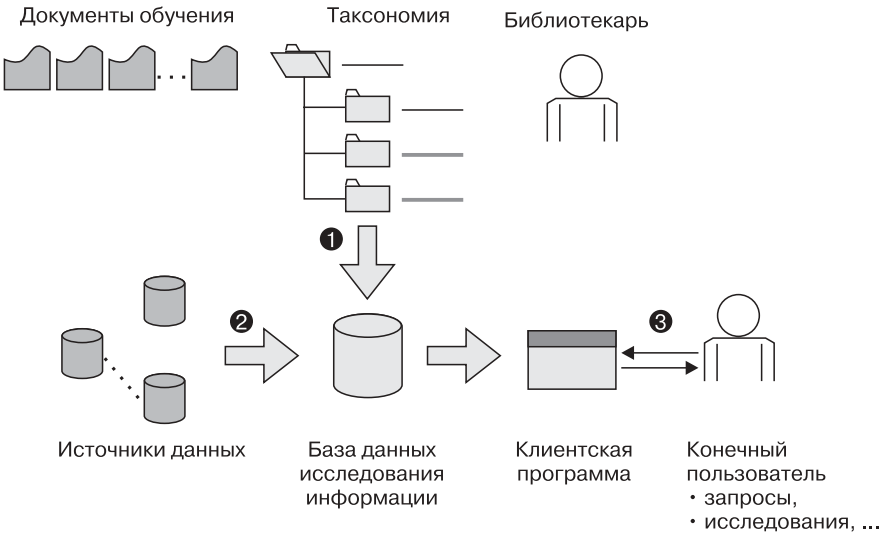

Рисунок 1. Задачи исследования информации

Чтобы использовать функций исследования информации, необходимо организовать документы для просмотра и перемещения по ним. Обычно эта задача выполняется библиотекарем или инженером по знаниям. Библиотекарь использует Information Structuring Tool (IST) для определения таксономии - иерархически организованных тематических характеристик данных документов, по которым в дальнейшем проводится анализ. IST - это программа с графическим пользовательским интерфейсом, позволяющая создавать и поддерживать таксономии. Для категорий выполняется обучение; получив стабильную таксономию, можно обращаться к различным источникам данных.

Документы могут импортироваться с любого контент-сервера или из Web с помощью службы Web Crawler (доступна только в виде JavaBeans) и распределяться по категориям.

К содержимому документа последовательно применяются функции текстового анализа и создания метаданных; они доступны в виде интерфейса программирования на уровне невизуальных JavaBeans и API служб Java.

JavaBeans исследования информации - это программные компоненты для быстрой разработки прикладных программ, соответствующие требованиям JavaBeans. API служб Java содержит все функции исследования информации в виде отдельных стандартных блоков для создания прикладных программ. Для помощи в создании прикладных программ поставляются примеры использования JavaBeans и JSP примера.

Для всех операций исследования информации, обрабатывающих документы, сначала надо выполнить распознавание содержимого документа. В эту задачу входят следующие подшаги:

1. Определение кодовой страницы документа
2. Определение обрабатываемых разделов текста, то есть пропуск информации разметки или двоичных данных (например, изображений)

Документы на контент-серверах могут иметь произвольную структуру, поэтому службы исследования информации позволяют создавать специальные модули, распознающие и выделяющие текстовые разделы из разных форматов документов. В службах исследования информации есть модуль по умолчанию, распознающий

множество часто используемых форматов документов. Список поддерживаемых форматов смотрите в разделе Глава 9, “Форматы документов”, на стр. 119, а подробное описание использования модуля по умолчанию смотрите в руководстве *Application Programming Guide*.

При создании метаданных для каждого из выбранных документов обрабатывается содержимое документа и применяются статистические методы и эвристические алгоритмы на основе ресурсов знаний, например, словарей и профилей частот

API исследования информации поддерживают следующие операции:
- Составление сводок
- Категоризация
- Идентификация языка
- Извлечение информации
- Кластеризация

Созданные метаданные для всех документов сохраняются на складе данных исследования информации.

После заполнения этого склада данных для выбора документов может использоваться еще один способ - выбор документов на основе информации из этого склада данных. В операции расширенного поиска используется сочетание текстового запроса и категории, что ограничивает область поиска документами из конкретной категории.

Использование исследования информации в деловой среде

В структуру организации, использующей реализацию какой-либо технологии исследования информации, обычно входят следующие должности:

- Системный администратор по информационным технологиям (его функции могут не ограничиваться только поддержкой исследования информации)
- Прикладной программист
- Библиотекарь или инженер по знаниям
- Люди, работающие с прикладной программой исследования информации (конечные пользователи)

В зависимости от природы прикладной программы могут потребоваться также следующие должности:

- Web-дизайнер
- Разработчик или консультант

Эти должности и действия исследования информации показаны на рис. 2 на стр. 45.

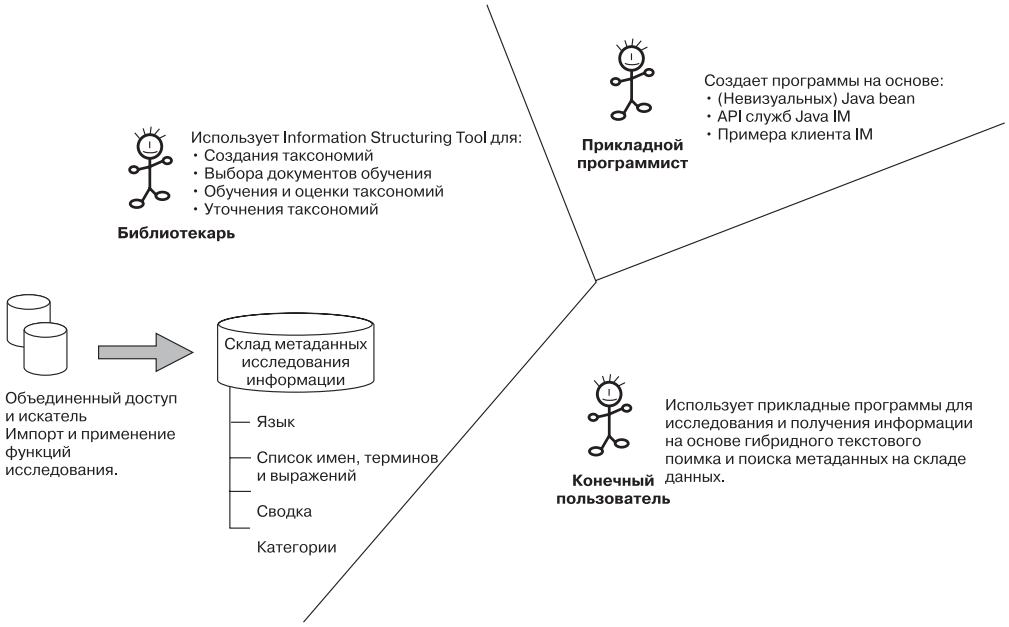

Рисунок 2. Должности и действия для исследования информации

Системный администратор настраивает аппаратную и программную среду и поддерживает требуемые ресурсы, например, пространство файловой системы и права доступа. Системный администратор устанавливает необходимые компоненты Enterprise Information Portal и конфигурирует источники содержимого и программу управления Enterprise Information Portal, обеспечивая доступ к различным хранилищам. Системный администратор также управляет складом метаданных исследования информации на уровне базы данных.

Прикладной программист создает прикладные программы, используя JavaBeans или API служб. Примеры JavaBeans смотрите в руководстве *Application Programming Guide for Windows*, а описание API служб - в руководстве *Application Programming Guide*. Прикладной программист может обращаться к помощи *Web-дизайнера*.

Библиотекарь или *инженер по знаниям* отвечает за настройку и поддержание собраний документов и ресурсов, используемых для исследования и получения информации. Библиотекарь использует программу управления Enterprise Information Portal для создания отображений метаданных и шаблонов поиска и программу Information Structuring Tool (смотрите раздел “Построение таксономии” на стр. 61) для определения каталогов и таксономий. Обычно библиотекарь отвечает также за заполнение склада данных исследования информации метаданными для документов из источников данных или из Web с помощью прикладных программ, написанных прикладным программистом.

Конечные пользователи работают с прикладными программами, созданными прикладным программистом, и выполняют задачи исследования и получения информации, используя ресурсы, которые созданы и поддерживаются библиотекарем или инженером по знаниям. В зависимости от распределения работы между конечными пользователями и библиотекарем, конечные пользователи могут также участвовать в выборе документов с контент-серверов и заполнении склада данных исследования информации.

Пример использования исследования информации

Фирма Electro Corp. производит электронные устройства для массового рынка. Она производит по крайней мере пять различных продуктов с широким диапазоном индивидуальных конфигураций.

В отделе сбыта есть информация о предпочтениях покупателей в определенных областях применения устройств. Профили использования показывают, как покупатели применяют эти продукты и какие существуют варианты их конфигурации. Каждый профиль относится к определенной стратегии маркетинга, выпуска продукции и управления связями.

В отделе обслуживания есть информация о комплектующих этих устройств, о сборке устройств из комплектующих, о поставщиках комплектующих и о надежности и ремонтопригодности отдельных деталей.

В отделе контрактов хранится информация о торговых посредниках и субподрядчиках. В нем также есть доступ к юридическим документам с условиями конкретных типов контрактов.

Недавно было отмечено падение объема продаж определенным покупателям. В результате конкуренции изменились методы применения разных электронных устройств и в соответствии с новыми достижениями технологии изменились ожидания покупателей.

Для адаптации к этим изменениям и использования их к своей выгоде в компании Electro Corp. создана рабочая группа стратегического планирования, которая позволит компании вернуться на свои позиции в бизнесе.

Первый шаг - разработать инфраструктуру информационной технологии для доступа к нужной информации, чтобы планировщики могли быстро принимать решения, владея полной информацией.

Примеры такой информации:

- Данные о продуктах конкурентов, их возможностях, ценах и приемлемости для покупателей
- Информация о сильных и слабых сторонах линии продуктов Electro Corps. с точки зрения покупателей
- Тенденции и перспективы в областях, в которых обычно используются эти продукты

Эта информация находится в разнородных источниках данных с разными аппаратными и программными платформами, по-разному организована (например, иерархически, с индексами или в плоском файле) и содержится в документах разных типов (например, в записях базы данных или в HTML-файлах).

На рис. 3 на стр. 47 показано использование исследования информации.

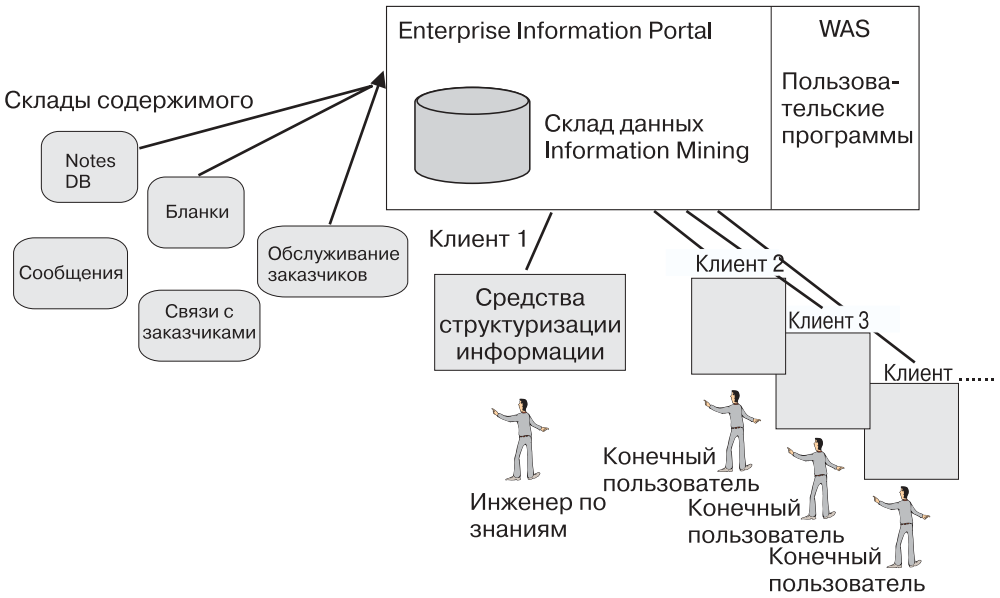

Рисунок 3. Пример исследования информации

Рабочая группа решает создать портал для доступа ко всем этим разнородным источникам из минимального клиента (в браузере), расположенного на клиентском компьютере планировщика. Выбирается Enterprise Information Portal, поскольку в нем есть все нужные блоки для создания такого портала и даже для создания специальных соединений с источниками юридических данных, созданным до появления стандартов в области деловых прикладных программ.

Эта задача включает в себя следующие шаги:

1. Создание аппаратной и программной инфраструктуры
2. Определение методов доступа к источникам данных, настройка необходимых соединений и создание отображений для нужных данных
3. Организация данных таким образом, чтобы планировщики могли просматривать их и перемещаться по ним
4. Создание прикладной программы для конечных пользователей

Исследование информации подключается на шаге 3. Когда инфраструктура доступна и работает, источники данных определены, соответствующие соединения установлены и нужные отображения созданы, к требуемым данным можно обращаться из одной точки и подмножества этих данных можно определить с помощью объединенного поиска. Следующий вопрос: как отфильтровать нужные данные для конкретного прогноза или выяснения тенденции и как организовать эти данные, чтобы их можно было использовать в процессе стратегического планирования.

Рабочая группа назначает *инженера по знаниям*, отвечающего за поддержание, организацию и обновление информации стратегического планирования. Для извлечения нужной информации из огромного множества документов, расположенных на разных источниках данных, инженер по знаниям опрашивает работников, которые участвовали в прошлых заданиях стратегического планирования, чтобы узнать взаимодействие процессов и полезные практические наработки, и находит сообщения от покупателей и базы данных поддержки.

С помощью средств поиска Enterprise Information Portal можно легко находить документы в этих базах данных по имени покупателя, адресу или характеристикам устройства. Однако информация, необходимая для определения профилей использования, скрыта в тексте, и для получения этой информации требуется интеллектуальный анализ содержимого документа с помощью служб исследования информации Enterprise Information Portal.

Полезный тип информации, создаваемой этими службами - тематическая характеристика содержимого документа, называемая категорией, например, *это документ о PDA*. Служба категоризации исследования информации анализирует содержимое документов и относит их к определенным категориям. Категории организованы в тематическую иерархию, называемую таксономией. Явно доступные и автоматически созданные метаданные располагаются в хранилищах (называемых каталогами), поддерживаемых этой службой исследования информации; эти хранилища помогают ускорить доступ к данным и их получение.

С помощью Information Structuring Tool инженер по знаниям определяет каталог, который показывает, как покупатели используют эти устройства.

Каталог показан на рис. 4 на стр. 48.

Рисунок 4. Пример каталога

Выполняя объединенный поиск по всем покупателям и способам использования устройств в базах данных продаж и поддержки, инженер по знаниям может определить набор учебных документов, представляющих каждую из категорий в таксономии. Это может привести к изменению таксономии, так как в новых категориях соответствующие данные могут оказаться более близкими.

Когда создана окончательная таксономия и для каждой категории есть достаточное число учебных документов, инженер по знаниям выполняет обучение таксономии с помощью Information Structuring Tool. При обучении создается модель категоризации, которую можно использовать для распределения документов по категориям с помощью службы категоризации.

Тем временем программисты из отдела информационных технологий создают прикладную программу для конечных пользователей - стратегических планировщиков, используя минимальный клиент Enterprise Information Portal и JavaBeans или API служб Java.

Эта прикладная программа состоит из набора шаблонов поиска, настроенных для стратегического планирования. С помощью этих шаблонов планировщики могут

заполнить каталог документами, полученными с разных серверов. При заполнении каталога служба исследования информации автоматически определяет категории для документов. Если нужно создать новый профиль использования, инженер по знаниям реорганизует таксономию в соответствии с новыми документами, указанными планировщиками в качестве учебного материала. Затем для каталога выполняется повторное обучение и новые результаты передаются планировщикам.

В изложенном примере показано, что используя функции исследования информации для отражения ожиданий и потребностей покупателей, такая компания, как Electro Corp., может идти в ногу с изменениями на рынке и сохранить свою конкурентоспособность.

Поддерживаемые языки и форматы

Службы исследования информации Enterprise Information Portal поддерживают следующие языки (смотрите Табл. 5):

Таблица 5. Поддерживаемые языки

Язык	Идентификация языка	Извлечение информации	Составление сводок	Категоризация	Кластеризация
Английский	x	x	x	x	x
Немецкий	x		x	x	x
Французский	x		x	x	x
Датский	x				
Финский	x				
Итальянский	x		x	x	x
Норвежский	x				
Португальский	x		x	x	x
Испанский	x		x	x	x
Шведский	x				
Корейский	x		x	x	x
Японский	x	x	x	x	x
Китайский (традиционный и упрощенный)	x		x	x	x

Список поддерживаемых форматов документов смотрите в разделе Глава 9, “Форматы документов”, на стр. 119.

Принципы

Объемы используемой информации постоянно растут. У большинства организаций есть большое и увеличивающееся число электронных документов, содержащих, возможно, очень существенную информацию, например, отклики покупателей, стратегическую информацию, жизненно важную для повышения конкурентоспособности на рынке, или информацию, позволяющую увидеть новые или изменившиеся благоприятные возможности для бизнеса. Службы исследования информации используются в роли программ, работающих с большими объемами электронных документов.

Архитектура системы

На рис. 5 показана архитектура системы исследования информации.

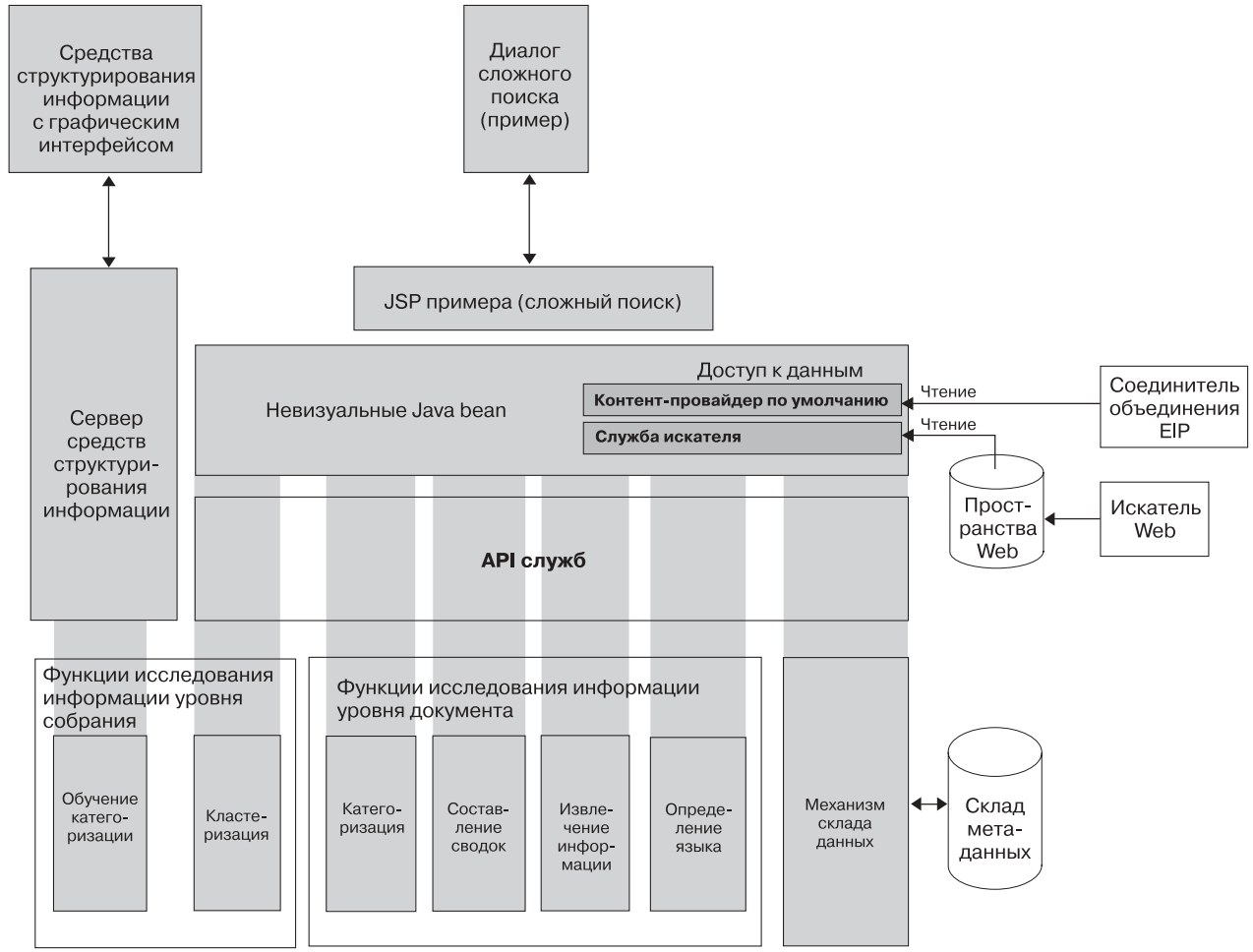

Рисунок 5. Архитектура системы исследования информации

Прямоугольники в правой части рисунка - это компоненты, используемые службами исследования информации, но не входящие в эти службы, а именно:

- Соединитель объединения (часть API OO)
- Web Crawler

Есть разные уровни использования функций исследования информации:

1. API служб Java

   Этот уровень представляет функции исследования информации и постоянных метаданных в виде согласованных API Java.

2. Невизуальные JavaBeans

   Этот уровень использует готовые компоненты, удовлетворяющие спецификациям JavaBeans и применяющие типы событий и соглашения из стандартных функций bean.

3. Примеры Java Server Pages

Этот уровень состоит из примера программы, использующей невизуальные JavaBeans и демонстрирующей, как создать прикладную программу для расширенного поиска, то есть текстового поиска с ограничением по категориям.

4. Information Structuring Tool

   Программа с графическим пользовательским интерфейсом для создания и поддержания таксономий.

Принципы исследования информации

В этом разделе описываются основные принципы исследования информации, что позволит вам хорошо понять и эффективно использовать функции исследования информации.

Службы исследования информации предоставляют инфраструктуру для создания и сохранения информации, относящейся к отдельным документам и собраниям документов. Такая информация о документе называется его **метаданными**.

Библиотека - это абстрактное представление содержимого базы данных исследования информации. Библиотека содержит набор каталогов.

Каталог - это склад метаданных для текстовых документов, содержащий:

- **Схему каталога**, определяющая, какие атрибуты сохраняются для каждого документа.
- **Таксономию** - иерархическую древовидную структуру **категорий**.
- **Модель категоризации**, построенную на основе обучения по документам; она может использоваться для автоматического назначения категорий для документов. Для генерации этой модели используется программа **Information Structuring Tool**, в которой можно создать таксономию и провести обучение. Эта модель используется службой категоризации.

Схема задает имена и типы атрибутов, которые могут генерироваться и сохраняться для документа в каталоге. Схема определена заранее и содержит следующие атрибуты:

- IKF_CONTENT типа string
- IKF_TITLE типа string
- IKF_AUTHOR типа string
- IKF_CATEGORIES типа string
- IKF_SUMMARY типа string
- IKF_LANGUAGE типа string
- IKF_FEATURES типа string
- IKF_COMMENTS типа string
- IKF_DATE типа отметка времени
- IKF_IDNUMBER типа integer

Для сохранения информации, которая извлечена из импортированного документа или создана на его основе, в каталоге создается **запись**, соответствующая схеме каталога. У записи есть уникальный идентификатор и набор пар значений имен. Уникальный идентификатор (который называют постоянным идентификатором объекта или PID) связывает создаваемую запись с исходным документом.

На рис. 6 на стр. 52 показан пример записи.

Рисунок 6. Пример записи

Если вы задаете значения для записи исследования информации при помощи JavaBeans или API служб, убедитесь, что размер используемых значений не превышают максимально допустимого. В противном случае вы получите `DKIKFSizeOutOfBoundsException`. Максимальные размеры:

Таблица 6.

Ключ	Максимальный размер в байтах
IKF_CONTENT	209715
IKF_TITLE	2048
IKF_AUTHOR	2048
IKF_CATEGORIES	8192
IKF_SUMMARY	8192
IKF_LANGUAGE	8
IKF_FEATURES	524288
IKF_COMMENTS	8192

Созданная запись сохраняется в каталоге в определенной категории. Для выбора категории обычно используются результаты категоризации, хотя категория может также выбираться по какому-либо другому значению из записи. Запись должна быть назначена в определенную категорию, поскольку при этом также происходит индексация содержимого документа и становится возможным текстовый поиск. У каждого каталога есть единый текстовый индекс, то есть все результаты поиска всегда автоматически оказываются в области поиска каталога.

Механизм склада данных - компонент, обеспечивающий доступ к складу постоянных данных.

Инструменты исследования информации

Службы исследования информации содержат функции для работы с электронными документами. В них входят:

- Программа Information Structuring Tool создает и поддерживает каталоги.
- Служба распознавания языка автоматически определяет язык, на котором написан документ.
- Служба категоризации автоматически относит документы к категориям, определенным ранее с помощью Information Structuring Tool.
- Служба составления сводок анализирует слова и фразы в документе и создает сводку для документа.

- Служба извлечения информации автоматически распознает важные элементы; при этом вам не нужно задавать словарь для конкретной предметной области.
- Служба кластеризации делит набор документов на группы или кластеры. Документы в одном кластере имеют общие свойства. Кластеры не определены заранее, они создаются автоматически.
- Функция расширенного поиска выполняет поиск текста в документах, сохраненных в каталоге; поиск ограничен определенными категориями.

Information Structuring Tool

Information Structuring Tool - это программа Web, позволяющая создавать и поддерживать набор каталогов, который называется библиотекой. Каталог используется для хранения метаданных, извлеченных из документов; он связан с таксономией, используемой для организации документов в виде заранее определенной структуры. Таксономия - это иерархическая структура категорий, используемая для классификации документов по тематике их содержимого.

Например, с помощью Information Structuring Tool библиотекарь может определить каталог о пищевых предпочтениях птиц.

На рис. 7 показан пример каталога.

Рисунок 7. Пример каталога

Хорошо структурированная система категорий значительно облегчает поиск нужной информации в большом объеме данных. Категории выбираются в соответствии с целями использования собрания документов; перед применением категорий выполняется обучение по примерам документов. Модель категоризации, созданная с помощью Information Structuring Tool, может затем использоваться службой категоризации для автоматического присваивания категорий документам.

В функции Information Structuring Tool входят:

- Создание, переименование и удаление каталогов.
- Редактирование описаний каталогов.
- Создание, переименование и удаление категорий.
- Добавление учебных документов к категориям и удаление их из категорий.

- Просмотр содержимого учебных документов.
- Запуск и остановка процесса обучения для каталога.
- Получение информации о качестве данных обучения в каталоге.

Подробное описание установки и использования Information Structuring Tool смотрите в разделе “Построение таксономии” на стр. 61.

Распознавание языка

Служба распознавания языка выбирает из данного набора языков тот язык, на котором, скорее всего, написан текстовый документ.

Служба распознавания языка для каждого документа возвращает упорядоченный список языков и значений достоверности. Язык указывается в виде двухсимвольного кода в соответствии со стандартом ISO 639. Значения достоверности указывают, насколько хорошо данный документ соответствует конкретному языку; эти значения выражены в виде чисел с плавающей точкой в диапазоне от 0 (плохо) до 1 (хорошо). Алгоритм распознавания языка предназначен для определения языка в документах, написанных на одном языке. Поэтому для документов, написанных на нескольких языках, нельзя гарантировать, что оценки по значениям достоверности будут правильно отражать язык документа.

Могут распознаваться следующие языки:
- Английский EN
- Немецкий DE
- Французский FR
- Датский DA
- Финский FI
- Итальянский IT
- Норвежский/букмол NB
- Норвежский/нюнорск NO
- Португальский PT
- Испанский ES
- Шведский SV
- Корейский KO
- Японский JA
- Упрощенный и традиционный китайский ZH

Можно задать следующее свойство:

- **maxResults** (только при использовании API служб Java):

  Максимальное число языков, определяемых и возвращаемых для каждого документа. Это целое значение, большее или равное нулю. Значение по умолчанию - 1, то есть возвращается один наиболее вероятный язык. Если задано значение 0, возвращаются все распознанные языки, для которых значение достоверности не ниже 0,01, в порядке значений достоверности.

Распознавание языка можно использовать в качестве шага предварительной обработки перед другими службами исследования информации. Например, можно использовать распознавание языка, чтобы перед извлечением информации найти все документы на английском или японском языке.

Категоризация

Категоризация - это отнесение документов к категориям и организация документов в соответствии с заранее определенной организационной моделью, созданной с помощью Information Structuring Tool.

Поэтому перед использованием службы категоризации необходимо с помощью Information Structuring Tool определить таксономию и выполнить ее обучение, чтобы создать такую модель.

Результат категоризации содержит категорию и значение достоверности, указывающее, насколько хорошо соответствует документ этой категории. Для каждого документа возвращается набор таких результатов. Их список упорядочен по значениям достоверности.

Можно задать следующие параметры:

- **maxResults**:

  Максимальное число категорий, возвращаемое для каждого документа. По умолчанию используется значение –1, при котором возвращаются все категории. Список результатов ранжируется.
- **minConfidence**:

  Назначенное документу значение достоверности указывает, насколько хорошо этот документ соответствует данной категории. Параметр minConfidence задает минимальное значение в диапазоне от 0 (самое плохое соответствие) до 1 (самое лучшее соответствие). По умолчанию используется значение 0, при котором возвращаются все назначенные для документа категории. Список результатов ранжируется от наилучшего соответствия к наихудшему.
- **catalogName**:

  Задает каталог, используемый для категоризации. Каталог может быть создан и обучен с помощью Information Structuring Tool.

Список поддерживаемых языков смотрите в “Поддерживаемые языки и форматы” на стр. 49.

Составление сводок

Сводка документа состоит из собрания извлеченных из документа фраз, характеризующих содержимое документа. Средство создания сводок может, например, помочь вам решить, интересен ли вам данный документ и нужно ли прочитать его целиком, или, если документ возвращен как часть результата запроса, нужно ли переходить по ссылке к этому документу.

Результат создания сводки содержит сводку в форме одной строки и в форме структуры данных (матрицы), которая может использоваться прикладными программами для выбора отдельных фраз и определения того, насколько близко они находятся.

Службу создания сводок можно использовать в разных режимах. Режим определяет, как будут использоваться значения `maxLength` и `ratio` для определения длины сводки.

- **maxLength**:

  Максимальное число предложений в сводке. Длина созданной сводки на будет превышать maxLength. По умолчанию используется значение 3.
- **ratio**:

  Число предложений относительно общей длины документа. Длина создаваемой сводки будет зависеть от общего числа предложений в документе. По умолчанию используется значение 0,1.
- **mode**:

  Определяет отношение между maxLength и ratio, необходимое для задания длины сводки. Используются различные режимы:

  – MODE_LESS_THAN_MAXLENGTH:

Число предложений сводки будет не больше числа maxLength. Это режим по умолчанию.

– MODE_EQUALS_RATIO:

  Число предложений в сводке определяется значением ratio. Оно вычисляется умножением ratio на общее число предложений в документе.

– MODE_EQUALS_RATIO_BUT_AT_MOST_MAXLENGTH:

  Число предложений сводки будет определяться значением ratio (ratio, умноженному на общее число предложений в документе), но не будет превышать maxLength.

Список поддерживаемых языков смотрите в “Поддерживаемые языки и форматы” на стр. 49.

Извлечение информации

Важная задача при анализе документов - извлечение элементов, дающих информацию о содержимом документа. Эти ключевые элементы могут использоваться:

- Как указатели важной информации, помогающие оценить интересность документа
- Для поиска и сохранения ключевых понятий, используемых для уточнения запроса
- Как критерий для сбора документов, связанных друг с другом

Примеры ключевых элементов - элементы словаря, такие как слова, имена и многосложные термины.

Для английского языка служба извлечения информации нормализует найденные ключевые элементы и группирует вхождения таких ключевых элементов, которые обозначают одно и то же или выражают одно понятие. Например, если в документе встречаются имена James J. Smith, Mr. Smith, James и Smith, они переводятся в нормализованную форму и отмечаются как имена одного и того же человека. Формы слов также переводятся в нормализованную форму, например, слово children переводится в child.

Однако для японского языка все ключевые элементы извлекаются в том виде, в котором они находятся в документе. Нормализация для них не выполняется (исключение - значения даты, времени и денежных единиц, нормализация которых соответствует ISO8601 и ISO4217).

Служба извлечения информации позволяет анализировать документы по:

- Односложным и многосложным элементам словаря, например, announcement, value, product cycle
- Географическим названиям, именам и названиям организаций, например, Washington, Bush, Data Management Academy
- Сокращениям, например, MB (мегабайты)
- Значениям даты, денежной величины и чисел, например, 11 Jan. 1958, 01/11/58, $30, thirty pence, 4.5, 5000

Можно задать три типа извлекаемой информации:

- Имя
- Термин
- Выражение

Используя API служб Java, можно также задать подтипы, перечисленные ниже, и значения достоверности, указывающие, насколько хорошо подтип соответствует

извлеченному элементу словаря. Значения достоверности находятся в диапазоне от 0 (плохо) до 1 (хорошо). Подтипы для типов имя, термин и выражение:

- **Имя**
  - Место, например, Montreal или London
  - Человек, например, Tim Brown
  - Организация, например, Smith and Son
  - Неизвестно, например, Smashing Pumpkins, Silicon Valley, CCTV (сокращение, для которого неизвестна полная форма)
  - Другое, например, AIS Plan, ISO Conference, Internet, Privacy Act Officer, JCAHO Performance Report
- **Термин**
  - Неустановленный термин, например, entertainment conglomerate, art world, class variable, source code, data definition, process improvement initiative
- **Выражение**
  - Количественное, например, four, fifty, 70
  - Порядковое, например, fourth, fiftieth
  - Процент, например, 12%, sixty percent
  - Дата, например, 07/28/98
  - Время, например, 18 hrs, 4 o'clock
  - Денежная, величина, например, DM90, thirty pounds
  - Сокращение, например, NY

Извлечение информации работает только для документов на английском или на японском языке. В качестве шага предварительной обработки можно использовать службу распознавания языка, чтобы определить, какие документы в собрании документов написаны на других языках. Службу извлечения информации можно комбинировать с другими функциями исследования информации, например, использовать ее в качестве шага предварительной обработки для службы составления сводок, чтобы создать сводки только для тех документов, в которых говорится о Буше - президенте, а не о Буше - губернаторе Техаса.

Кластеризация

Средство кластеризации организует собрание документов таким образом, чтобы схожие документы были в одной группе, а документы в разных группах (кластерах) отличались по содержанию. Таким образом, кластеризация может использоваться как средство, дающее общее представление о большом собрании документов и находящее документы, связанные друг с другом. Кластеризацию можно также использовать для поддержки построения таксономии с помощью Information Structuring Tool, кластеризуя документы для обучения внутри прикладной области. Кластеризация может быть также полезна для поиска в собрании сходных документов, которые могут указывать на новые тенденции или новые технологии, и для нахождения аналогичных или очень похожих документов, которые могут быть интересны для анализа конкурентной среды.

Кластеризация - это итеративный процесс, организующий документы в кластеры так, чтобы документы в каждом кластере были как можно более похожими (по содержимому документов), а кластеры как можно сильнее отличались друг от друга. Кластеризация работает с собранием документов как целым, в противоположность описанным выше службам исследования информации, таким как категоризация и составление сводок, которые работают на уровне документа. При кластеризации

типичные характеристики каждого документа сравниваются с характеристиками других документов и документы объединяются в группы в соответствии с их сходством.

На фазе кластеризации нельзя добавлять новые документы в набор документов.

Можно задать следующие параметры:

- **maxClusterCount**

  Максимальное число возвращаемых кластеров.
- **minClusterCount**

  Минимальное число возвращаемых кластеров.
- **clusterFeatureCount**

  Число меток (ключевых слов), возвращаемых для каждого кластера.

Однако эти значения являются жесткими условиями для программы кластеризации, они лишь представляют собой рекомендованные границы. Служба кластеризации возвращает список результатов.

Кластеризация работает только для документов на английском языке. В качестве шага предварительной обработки можно использовать службу распознавания языка, чтобы определить, какие документы в собрании документов написаны на других языках.

Расширенный поиск

В отличие от обычного поиска Enterprise Information Portal, который работает со всем содержимым контент-сервера Enterprise Information Portal, так называемый расширенный поиск выполняет поиск только среди документов, ID которых хранятся в каталоге, созданном с помощью Information Structuring Tool. Чтобы еще более уточнить область поиска, в запросе расширенного поиска можно не только задать текст для поиска, но и ограничить область поиска документов отдельными категориями.

Можно задать следующие параметры:

- **catalogName**:

  Задает каталог, используемый для поиска. Каталог может быть создан и обучен с помощью Information Structuring Tool.
- **maxResults**:

  Максимальное число результатов поиска, возвращаемых для каждого запроса. Значение по умолчанию - 0 (возвращаются все результаты).

Можно использовать следующие типы запросов:

1. Чисто текстовый запрос. Такой поиск возвращает все документы, удовлетворяющие текстовому критерию. Список результатов упорядочен по релевантности.
2. Чистый поиск по категории. Такой поиск возвращает все документы, которым назначена заданная категория. Результаты возвращаются в произвольном порядке.
3. Комбинированный поиск по тексту и категории. Такой поиск возвращает все документы, удовлетворяющие текстовому критерию и относящиеся к заданной категории. Список результатов упорядочен по релевантности.

Запросы расширенного поиска всегда связаны с конкретным каталогом. Говорят, что областью поиска такого запроса является каталог. Поиск сразу в нескольких

каталогах невозможен, поскольку каталоги - это способ представления импортированных документов, который не должен нарушаться.

Синтаксис запроса (бэкусовская нормальная форма) для строки запроса:

```
строка_запроса       ::= терм
терм                 ::= ( терм )
                     ::= простой_терм
                     ::= составной_терм
простой_терм         ::= терм_категории
                     ::= терм_текстового_поиска
                     ::= строчный_терм
                     ::= числовой_терм
составной_терм       ::= терм бинарная_логическая_операция терм
                     ::= унарная_логическая_операция простой_терм
терм_категории       ::= ( DKIKFCategory операция_категории путь_категории )
терм_текстового_поиска ::= ( "имя атрибута" CONTAINS значение_текстового_поиска )
строчный_терм        ::= ( "имя атрибута" строчная_операция строчное_значение )
числовой_терм        ::= ( "имя атрибута" базовая_операция числовое_значение )
бинарная_логическая_операция ::= AND | OR
унарная_логическая_операция  ::= NOT
операция_категории   ::= >= | =
строчная_операция    ::= LIKE | базовая_операция
базовая_операция     ::= > | < | <= | >= | != | =
путь_категории       ::= "путь категории"
значение_текстового_поиска   ::= "'строка'"
строчное_значение    ::= "'строка'"
числовое_значение    ::= "целое число" | "десятичное число"
```

- Разделители, строки и числа представляют собой общие термы.
- Операция категории '=' ограничивает область поиска только одной категорией.
- Операция категории '>=' задает в качестве области поиска эту категорию и все ее подкатегории в дереве категорий.
- Строка поиска в условии `CONTAINS` может включать символы подстановки: `('_')` для одного символа и `('%')` для произвольного числа символов. Например, `_LOB` может соответствовать `BLOB` и `CLOB`, а `%name` может соответствовать `filename`. С помощью строчной операции `CONTAINS` можно выполнять поиск только по тем атрибутам схемы, которые отмечены как атрибуты с возможностью поиска, например, IKF_CONTENT.
- Строка поиска в условии `LIKE` может также содержать символы подстановки, применяемые в SQL.
- Полный список поддерживаемых в настоящее время имен атрибутов смотрите в разделе “Принципы исследования информации” на стр. 51.

Примеры запросов:

- Чисто текстовые запросы:

  ```
  ("IKF_CONTENT" CONTAINS "'southern Africa'") AND NOT
  ("IKF_CONTENT" CONTAINS "'Cape'")
  ```

- Чистые запросы по категориям:

  ```
  ("DKIKFCATEGORY" >= "birds/Fruit eaters")
  ```

- Комбинированный запрос по тексту и категории:

  ```
  ("IKF_CONTENT" CONTAINS "'South Africa'") AND
  ("DKIKFCATEGORY" >= "birds/Birds of prey and scavangers/Falcon")
  ```

- Запрос по атрибуту:

  ```
  ("IKF_SUMMARY" LIKE "humming birds in the tropics")
  ```

  или

  ```
  ("IKF_FEATURES" LIKE "Goethe") AND ("IKF_TITLE" = "Faust")
  ```

Интерфейсы программирования

Для построения прикладных программ функции исследования информации доступны в виде:

- API служб Java
- JavaBeans исследования информации

API служб Java включают все функции исследования информации, за исключением поддержки каталогов, входящей в Information Structuring Tool как служба Enterprise Information Portal. Этот API обеспечивает связь клиент-сервер на основе RMI Java.

Используя API служб Java, прикладная программа может:

- Определить язык, на котором написан документ
- Создать сводки для текстовых документов
- Назначить категории документам
- Извлечь информацию (например, имена, термины или выражения) из текстового документа
- Объединить схожие документы в группы
- Сохранять и находить в каталоге метаданные для документов
- Выполнять текстовый поиск по документам, ограниченным определенными категориями, и по атрибутам, например, по сводке

API служб Java могут выполняться в локальном режиме (используя прямой вызов метода) или в удаленном режиме (используя вызов удаленного метода - RMI). Режим удаленного выполнения позволяет сконфигурировать один сервер в качестве сервера программ, на котором выполняются ваши программы Web, и другой сервер в качестве сервера исследования информации, выполняющего текстовый анализ, индексацию и поиск. Выполненные задачи посылаются механизмом задач сервера на сервер исследования информации (удаленный компьютер), и вся обработка выполняется на этом компьютере.

Если используется Web Crawler, для реализации соответствующих механизмов доступа надо применять JavaBeans. Для доступа к Web Crawler нельзя использовать уровень API служб Java.

Подробное описание API служб Java исследования информации смотрите в книге *Application Programming Guide for Windows*.

На рис. 8 на стр. 60 показана удаленная конфигурация исследования информации.

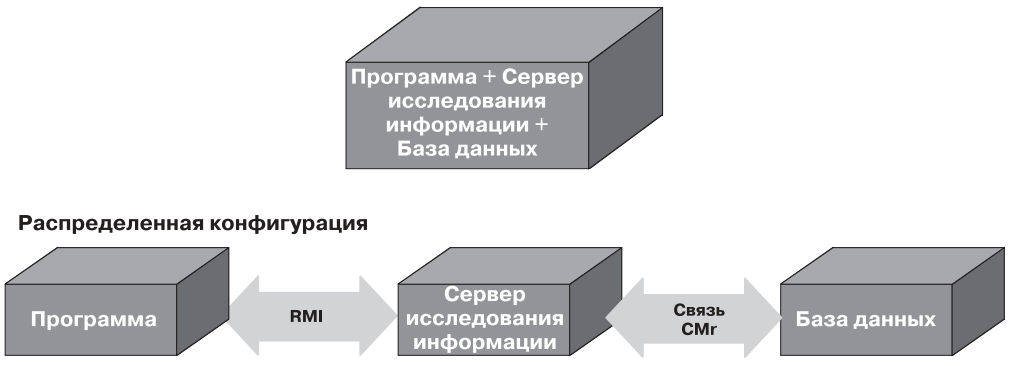

Рисунок 8. Удаленная конфигурация исследования информации

Функции bean исследования информации - это API Java высокого уровня, предназначенный для быстрой разработки прикладных программ и построенный в соответствии со спецификацией JavaBeans. Эти функции bean не поддерживают обработку задач сервера, поэтому для хорошей производительности следует выполнять всю разработку прикладной программы, использующей эти функции bean, на одном компьютере.

Каждая из этих функций bean может управляться событиями, а у некоторых из них есть методы для прямого вызова. Для интеграции с существующими функциями bean Enterprise Information Portal применяется поддержка события результата, используемого функциями bean Enterprise Information Portal. Это означает, что события результатов объединенного поиска и Web Crawler совместимы со службами исследования информации, и результаты исследования информации могут обрабатываться с помощью EIP. Подробное описание функциями bean исследования информации смотрите в книге *Application Programming Guide for Windows*.

Первые шаги

Первые шаги исследования информации Enterprise Information Portal - сценарный учебник, на практических примерах демонстрирующий инженерам по знаниям, администраторам и прикладным программистам, как можно применять технологию исследования информации IBM в деловой среде. Структура этого учебника:

- Краткое введение
- Что нужно сделать перед запуском Первых шагов
- Организация данных примера
- Доступ к этим данным
- Использование клиента примера
- Удаление данных примера
- Список литературы

Чтобы вызвать Первые шаги исследования информации, запустите файл `<CMBROOT>\ikf\firststeps\first_steps.html`.

Чтобы запустить Первые шаги исследования информации в Windows, Нажмите кнопку Пуск и выберите **Enterprise Information Portal for Multiplatforms 8.1 → <Исследование информации> → Первые шаги** .

Первые шаги исследования информации Enterprise Information Portal используются также для проверки установки.

Построение таксономии

Information Structuring Tool - это программа Web, позволяющая создавать и поддерживать набор каталогов, который называется библиотекой. Каталог используется для хранения метаданных; он связан с таксономией, используемой для организации информации в каталоге. Таксономия - это иерархическая структура категорий для классификации документов по тематике их содержимого. Например, верхний уровень таксономии может содержать такие категории, как *бизнес, культура, спорт*, в то время как на следующем более низком уровне *спорт* подразделяется на *игровые виды спорта* и *атлетику*. На еще более низком уровне *игровые виды спорта* (командный спорт) подразделяется на *футбол, хоккей, теннис*.

Назначая учебные документы для категорий и обучая каталог, Information Structuring Tool создает модель категоризации, которая используется затем службой категоризации для назначения категорий для документов.

Установка Information Structuring Tool

Information Structuring Tool должен размещаться в виде программы Web в контейнере сервлетов, например, в механизме сервлетов IBM WebSphere Application Server (WAS).

Однако учтите, что две программы Web Information Structuring Tool (например, одна с именем IST1 и другая с именем IST2) не могут одновременно работать с одним экземпляром исследования информации.

Перед внедрением Information Structuring Tool убедитесь, что установлена и сконфигурирована правильная версия WAS. Подробную информацию смотрите в книге*Планирование и установка Enterprise Information Portal*.

Права пользовательского доступа, необходимые для внедрения Information Structuring Tool:

- Для Windows: полномочия администратора
- Для AIX: привилегии пользователя root
- Для Sun: привилегии пользователя root

Начинаем работу

У этой программы есть Web-интерфейс, при помощи которого можно определять, поддерживать и обучать таксономии. Окно программы содержит две панели. Левая называется панелью *каталогов*; она используется для создания и поддержания таксономий. Правая панель называется *записной книжкой*; в ней выводится информация об использовании этой программы. Записная книжка содержит набор вкладок для обучения и оценки документов для каталога. Чтобы использовать программу, нужно сначала создать на левой панели каталогов каталоги и категории.

Права доступа

В Information Structuring Tool имя пользователя и пароль сохраняются системой Enterprise Information Portal. Вы можете ввести только имя пользователя - пароль известен Enterprise Information Portal.

При запуске Information Structuring Tool появляется предупреждение защиты, так как апплету Java, который используется для выгрузки на сервер учебных документов, требуется доступ для чтения к вашей файловой системе. Если вы откажетесь, вы не сможете выгружать учебные документы и, следовательно, не сможете обучать таксономии.

Information Structuring Tool может работать в многопользовательской среде. Чтобы разрешить нескольким пользователям просматривать одну таксономию, Information Structuring Tool поддерживает механизм блокировки, управляющий доступом к каталогу и его категориям.

Пользователь может заблокировать каталог явно, выбрав и заблокировав каталог перед началом работы с ним, или же начать работу с каталогом, например, добавлять учебные документы, в этом случае каталог блокируется автоматически для предотвращения конфликтов доступа. Другие пользователи могут просматривать этот каталог, но не могут ничего в нем изменить, пока каталог не будет разблокирован пользователем, заблокировавшим этот каталог.

Учтите, что если сервер программ, где размещен Information Structuring Tool, завершил работу, все блокировки удаляются.

Определение таксономии

Таксономия, представляющая собой древовидную структуру из категорий, привязана к определенному каталогу.

Шаги определения нового каталога и выбора соответствующих категорий:

1. Определите, какие категории нужно создать, и создайте новую таксономию.

   В панели каталогов выберите **Библиотека** и щелкните правой кнопкой мыши. Появится меню. Выберите **Новый каталог** - будет создан значок каталога. Переименуйте этот значок, введя имя каталога, и нажмите клавишу **Enter**. Будет создана папка с этим именем. Это корневая категория. Содержимое записной книжки изменится.

   Другой способ добавления нового каталога в библиотеку - импортировать существующую таксономию, созданную вне Information Structuring Tool, например, в файловой системе. Дополнительную информацию смотрите в разделе “Выгрузка учебных документов” на стр. 65.

   Чтобы вы могли активно работать с таксономией, каталог должен быть заблокирован вами. Когда вы создаете каталог, он автоматически блокируется. Чтобы заблокировать существующий каталог, выберите его и щелкните правой кнопкой мыши. Появится меню. Выберите **Заблокировать каталог** - значок состояния каталога изменится. Значки, используемые для различных типов состояния каталога:

   

   Дерево таксономии свернуто и не заблокировано ни одним пользователем.

   

   Дерево таксономии развернуто и не заблокировано ни одним пользователем.

   

   Дерево таксономии свернуто и заблокировано текущим пользователем.

   

   Дерево таксономии развернуто и заблокировано текущим пользователем.

   

   Дерево таксономии свернуто и заблокировано другим пользователем.

   

   Дерево таксономии развернуто и заблокировано другим пользователем.

   На рис. 9 на стр. 64 показан пример двух каталогов.

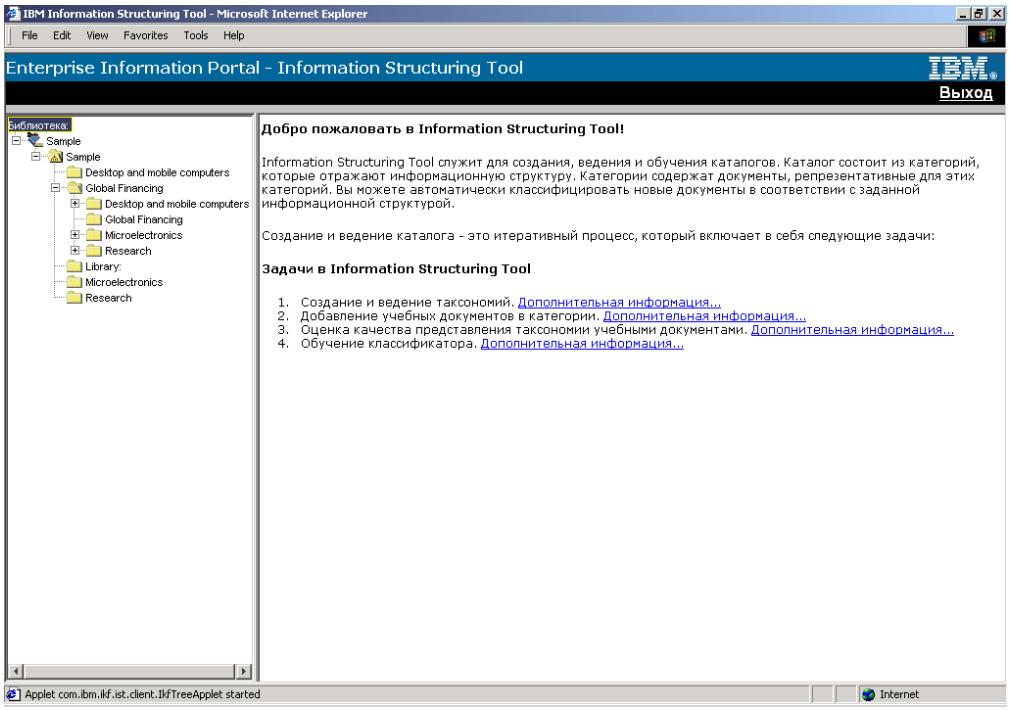

Рисунок 9. Пример каталога

2. Когда каталог создан и заблокирован, можно добавлять, переименовывать или удалять категории. При создании нового каталога создается также корневая категория, имя которой совпадает с именем каталога. Эту категорию можно переименовать. Для этого выберите категорию и щелкните правой кнопкой мыши. Выберите **Переименовать**, введите новое имя и нажмите клавишу **Enter**.

   Чтобы добавить новые категории, выделите категорию, в которую нужно добавить новую подкатегорию, и щелкните правой кнопкой мыши. Выберите **Новая категория**, введите новое имя и нажмите клавишу **Enter**. Имена категорий на одном уровне древовидной структуры должны быть уникальными.

   Корневую категорию нельзя удалить. Она удаляется только при удалении каталога. При удалении категории удаляются также все ее подкатегории (категории, расположенные ниже в дереве категорий), все учебные документы и все записи в этой категории.
3. Информация описания для выбранного каталога выводится на вкладке **Свойства**. Изначально поле описания пусто. Чтобы добавить или отредактировать описание каталога, нажмите кнопку **Редактировать описание**. Откроется окно, в котором можно ввести описание.

Выбор учебных документов

Качество модели категоризации сильно зависит от того, насколько правильно заданы учебные документы для каждой категории.

Учебные документы должны быть в одном из форматов документов, поддерживаемых Enterprise Information Portal. Список поддерживаемых форматов смотрите в разделе Глава 9, “Форматы документов”, на стр. 119.

Очень важно выбрать подходящий набор учебных документов. Документы должны:

- Быть типичными представителями категории

- Содержать значительный объем описательного текста, в котором не слишком велика доля разметки и списков слов.
- Быть написанными в одном стиле, например, нужно избегать документов в стиле описания, если остальные документы написаны в стиле отчета.
- Быть примерно одной длины и, желательно, не слишком длинными; учебные документы должны также быть примерной той же длины, что и документы, для категоризации которых будет использоваться служба категоризации.

Рекомендуется выбирать примерно 40 учебных документов для каждой категории; однако для более общей категории требуется больше документов. Выбирайте осмысленные категории; если вы сами не очень четко понимаете категории и учебные документы и не можете проводить классификацию по ним, то и автоматическая обработка не даст хороших результатов.

Выгрузка учебных документов

Чтобы добавить учебные документы, выберите соответствующую категорию - появится окно **Список учебных документов**. Чтобы добавить документы в этот список, нажмите кнопку **Добавить документ**.

Появится окно **Добавление учебных документов**. Не обязательно закрывать это окно после выгрузки файлов - его можно использовать для последующих операций выгрузки файлов для другой категории или другого каталога.

Чтобы добавить учебные документы, нажмите кнопку **Обзор** и выберите нужные файлы или каталог файловой системы в окне **Открыть**.

Для выгрузки можно выбрать один или несколько файлов из одного каталога файлов или каталог целиком. Если этот каталог пуст, будет выдано сообщение об этом.

Это позволяет импортировать и использовать существующие таксономии, созданные вне Information Structuring Tool, например, в файловой системе.

Например, если в файловой системе выбран каталог Development, содержащий подкаталог Design, а категория Development в дереве таксономии также содержит подкатегорию Design, файлы из этой подпапки будут добавлены в эту подкатегорию. Если эта подкатегория не существует, она будет создана и файлы будут добавлены в эту созданную подкатегорию.

Выберите язык документа и формат для всех выбранных файлов. За исключением плоских текстовых файлов, всегда используйте формат ″автоматическое определение″.

Чтобы добавить эти файлы в список учебных документов, нажмите кнопку **Передать**.

На рис. 10 на стр. 66 показано добавление учебных документов.

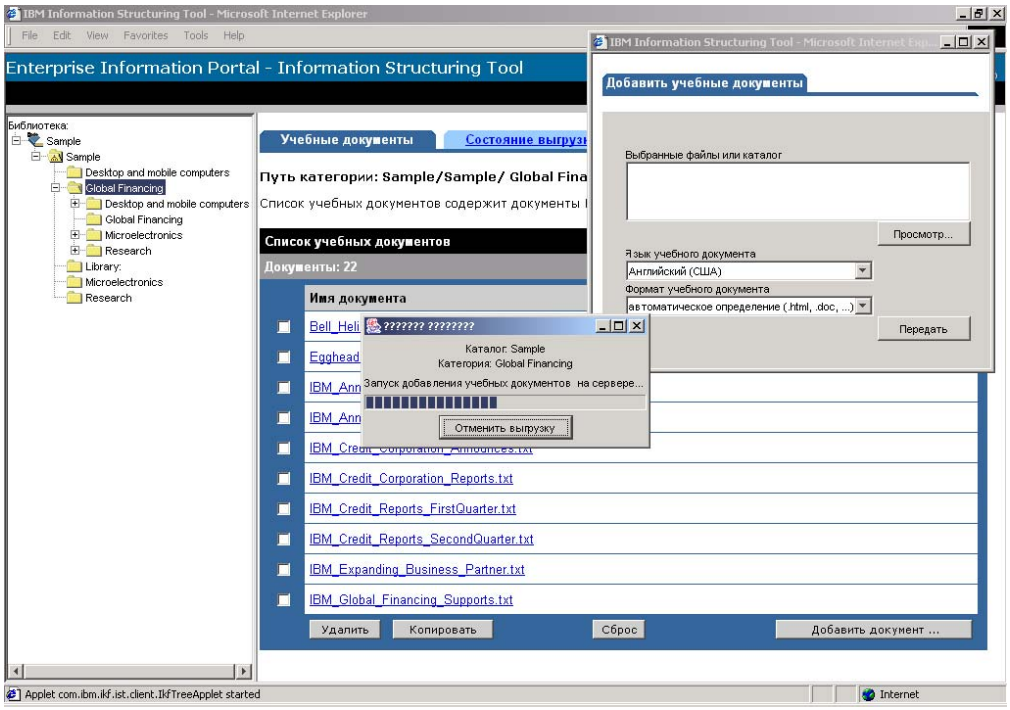

Рисунок 10. Добавление учебных документов

Появится окно **Процесс выгрузки**. В нем можно отменить процесс выгрузки документов. Если процесс выгрузки отменен, остальные файлы не добавляются. Уже добавленные файлы учебных документов удаляются. Однако если при выгрузке были добавлены подкатегории, они не удаляются из каталога.

Если файлы не удалось успешно выгрузить, автоматически выводится окно **Состояние выгрузки**, в котором указывается причина неудачи. Например:

- Файл с тем же именем уже существует.
- Файл пуст.
- Файл не удалось выгрузить на сервер.

В окне Состояние выгрузки щелкните по вкладке **Учебные документы**, чтобы вернуться в список учебных документов. Здесь выводятся все успешно выгруженные в категорию учебные документы. Чтобы увидеть все документы в списке, используйте кнопки **Предыдущий** и **Далее**.

Если один документ нужно выгрузить в качестве учебного документа в несколько категорий, выгрузите его в первую категорию и затем скопируйте в другие категории. Не нужно еще раз выгружать этот файл. Для копирования выберите один или несколько документов в окне Список учебных документов и нажмите кнопку **Копировать**. В появившемся окне нажмите кнопку **Обзор...**, чтобы выбрать одну или несколько к категорий, в которые нужно скопировать документы, и нажмите кнопку **Передать**.

Во время выгрузки файлов не разрешено выполнять следующие действия:

- Разблокирование каталога
- Выход из Information Structuring Tool
- Запуск обучения или оценки каталога
- Переименование каталога

- Удаление информации состояния выгрузки из каталога

Однако разрешены следующие действия:

- Запуск другого процесса выгрузки файлов (в ту же категорию или в другую)
- Работа с другим каталогом

На рис. 11 показан список документов для обучения.

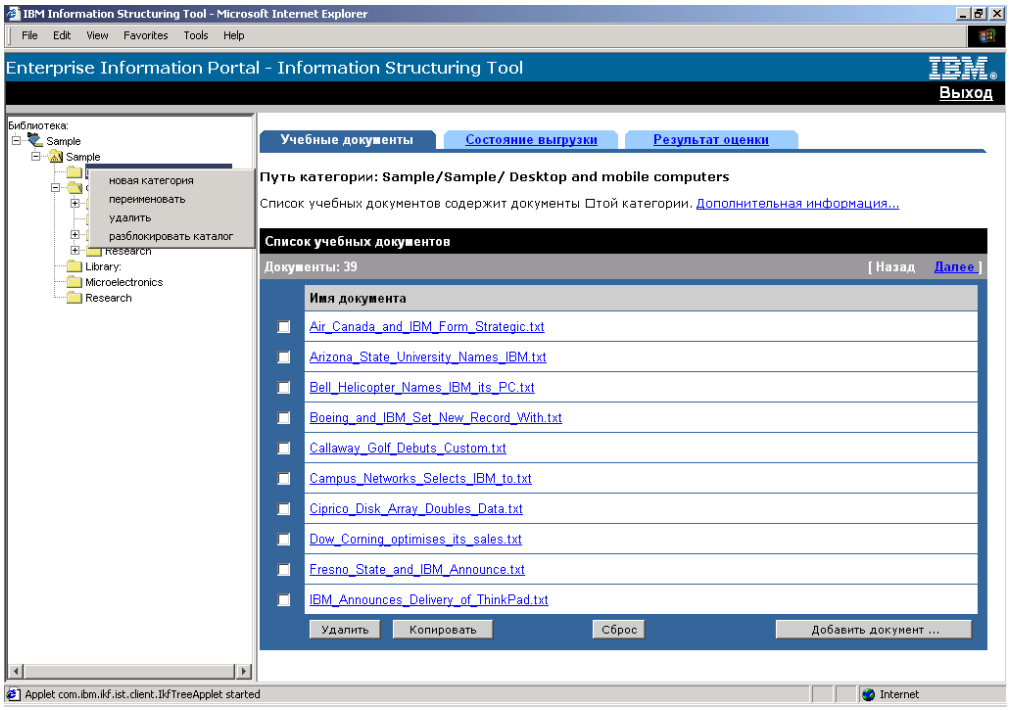

Рисунок 11. Список учебных документов

Если браузер был закрыт во время выгрузки файлов, файлы будут добавлены в качестве учебных документов, если они уже переданы на сервер. Если это не так, файлы не будут добавлены.

Выгрузив все учебные документы, выберите каталог, чтобы начать оценку.

Оценка модели категоризации

Определив таксономию и назначив учебные документы для каждой категории, нужно выполнить оценку таксономии. Оценка таксономии помогает узнать, насколько хорошо заданные учебные документы соответствуют данной таксономии. Это итеративный процесс, состоящий из следующих шагов:

1. Запуск оценки
2. Обращение к результатам оценки
3. Изменение таксономии или набора учебных документов
4. Повторное выполнение оценки

На каждой итерации оценки процесс оценки:

- Делит учебные документы на два набора: набор обучения (примерно 80% документов) и набор тестирования (примерно 20% документов). Выполняет обучение каталога, используя набор обучения, и применяет службу категоризации к набору тестирования.

- Проверяет, были ли документам назначены правильные категории с достаточно большими значениями достоверности. Значения достоверности лежат в диапазоне от 0 до 1, где 1 означает идеальное соответствие документа. Значение достоверности можно задать. Значение по умолчанию - 0,5.

Можно выбрать число итераций (от трех до пяти). По умолчанию используются три итерации, это позволяет хорошо увидеть сильные и слабые стороны таксономии. Если выбрать пять итераций, все документы будут входить как в набор обучения, так и в набор тестирования.

Чтобы начать процесс оценки, нажмите кнопку **Запустить оценку**.

На каждой итерации оценки для каждой категории вычисляются:

- Число *правильных* документов. Число учебных документов в этой категории, отнесенных в эту же категорию при оценке.
- Число *исключаемых* документов. Число учебных документов в этой категории, отнесенных при оценке в другую категорию.
- Число *включаемых* документов. Число учебных документов, отнесенных при оценке в эту категорию, но изначально назначенных другой категории.
- Число *нераспределенных* документов. Число учебных документов в этой категории, которые при оценке не были отнесены ни к одной из категорий. Сюда могут входить документы, которые были отнесены к какой-то категории, но значение достоверности оказалось меньше заданного.

На экран выводится информация о ходе процесса оценки на уровне каталога:

- Состояние процесса оценки, указывающее, выполняется ли процесс оценки или остановлен
- Дата последней оценки
- Число выполненных итераций оценки
- Общее число итераций
- Усредненная общая точность, показывающая процентную долю правильных документов, то есть тех документов изначально назначенных для этой категории, а также включаемых документов
- Усредненный общий выход, показывающий процентную долю правильных документов в этой категории
- Правильные документы, то есть число правильно назначенных документов
- Неправильно классифицированные документы, то есть число исключаемых и включаемых документов
- Нераспределенные документы

Значения точности и выхода тесно связаны с заданным значением достоверности. Если задано малое значение достоверности, значение точности понижается, а значение выхода растет, и наоборот. Высокая точность означает, что из назначенных документов многие учебные документы были назначены правильно; высокий выход, в свою очередь, означает, что большая часть учебных документов были назначены какой-то категории или, другими словами, нет или есть лишь немного нераспределенных документов.

Результаты оценки

Чтобы увидеть подробные результаты оценки, щелкните по вкладке **Результат оценки**. Если выбрать каталог и затем эту вкладку, будут показаны результаты для всего каталога; если выбрать категорию, будут показаны результаты для этой категории.

На рис. 12 показаны результаты оценки на уровне каталога:

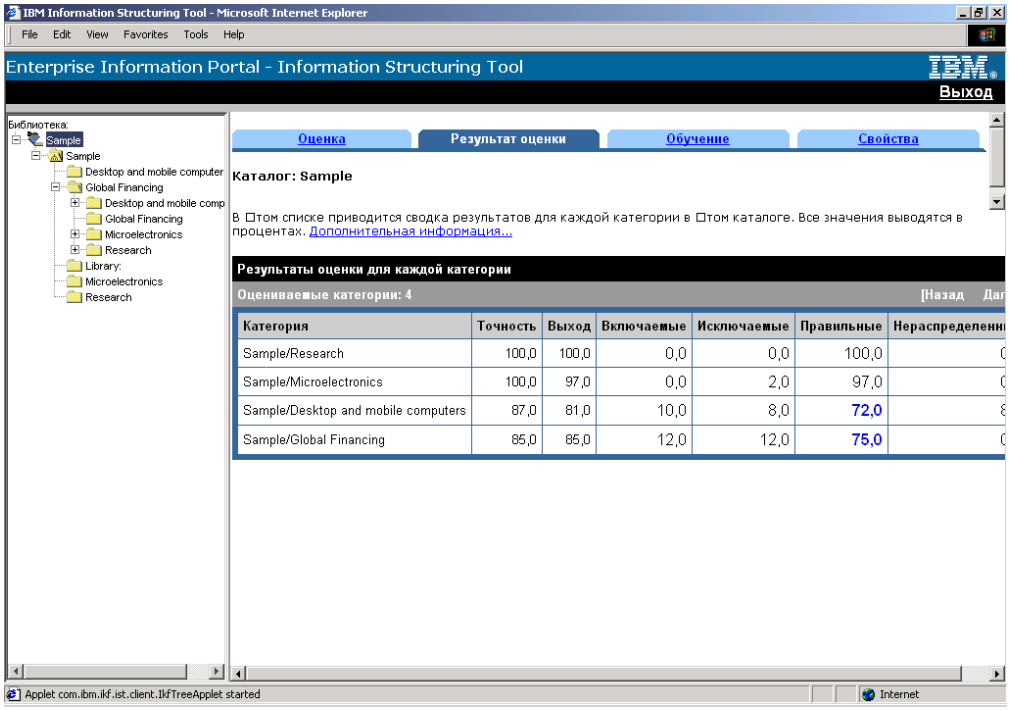

Рисунок 12. Результаты оценки на уровне каталога

Начните с общих результатов для каталога. Значения, выведенные красным цветом (самые важные) или синим цветом (менее важные), указывают на то, что что-то не так: с категорией или учебными документами. Качество категории определяется по ее учебным документам, поэтому целесообразно изучить перемещения документов в категорию (включаемые) и из категории (исключаемые).

Изменения, вносимые в категорию или в учебные документы, сильно зависят от того, придается ли основное значение точности и выходу в равной мере, или же только точности. Чем выше значение точности, тем более различительна данная категория по отношению к другим категориям таксономии. С другой стороны, чем больше значение выхода, тем меньше будет нераспределенных учебных документов.

Результаты оценки выводятся на двух уровнях:

1. Уровень **каталога**:

   Для каждой категории в это каталоге:

   - Процентные значения точности и выхода
   - Процентная доля включаемых документов
   - Процентная доля исключаемых документов
   - Процентная доля правильных документов
   - Процентная доля нераспределенных документов

2. Уровень **категории**:

   Для исключаемых и включаемых документов:

   - Учебные документы и исходные или нужные категории

   Для правильных и нераспределенных документов:

   - Учебные документы

Интерпретация результатов оценки

В следующем разделе показано, как можно интерпретировать результаты оценки; однако не забывайте, что таксономия действует как одно целое, то есть изменения в одном разделе таксономии могут негативно влиять на результаты, полученные в каком-то другом месте таксономии.

На рис. 13 показаны результаты оценки на уровне категории:

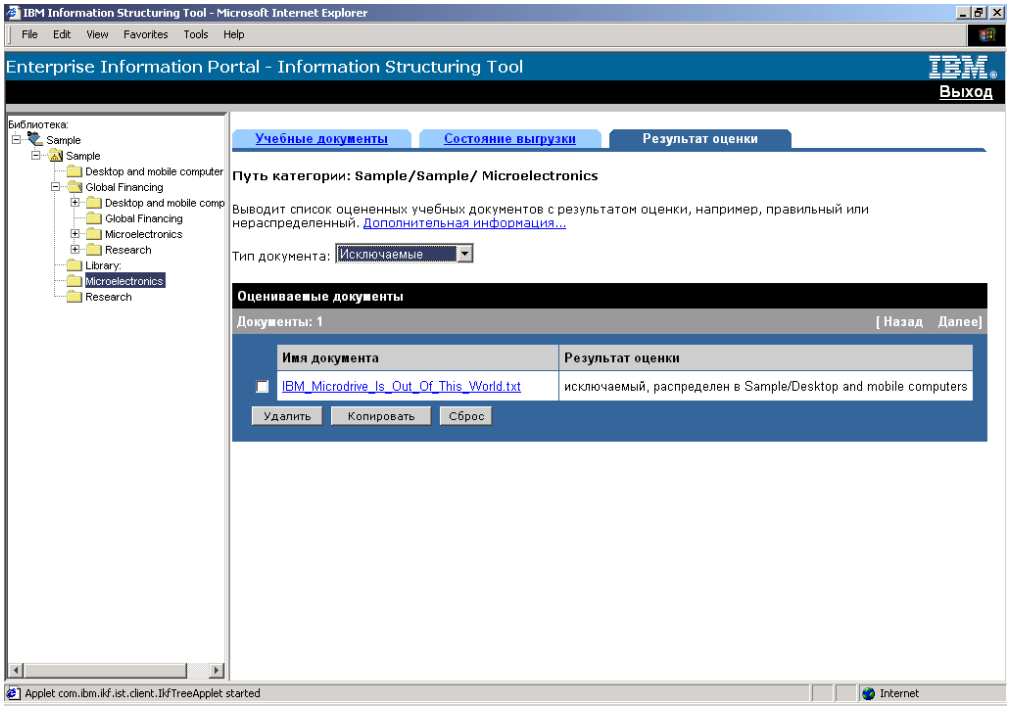

Рисунок 13. Результаты оценки на уровне категории

Начав с результатов оценки на уровне каталога, выберите категории с потенциально низкими значениями точности и выхода и со значениями учебных документов, отмеченными красным (самые важные) или синим (менее важные). Для каждой категории (категория выбирается в левой панели), в свою очередь, проверьте следующее:

- Много **включаемых** документов:
  - Категория получает документы из многих других категорий:
    - Категория недостаточно различительна. Скопируйте эти документы в эту категорию.
    - Сделайте эту категорию различительной, выбрав более подходящие учебные документы или разделив категорию на подкатегории.
  - Категория получает документы из одной или двух категорий:
    - Скопируйте эти документы в эту категорию.
    - Проверьте исходную категорию. Есть ли смысл сохранять эту категорию, или категории лучше объединить? Чтобы объединить категории, скопируйте все учебные документы в категорию, которую вы хотите сохранить, и удалите другую категорию.
- Много **исключаемых** документов:
  - Документы из этой категории попадают во много других категорий:

- Категория недостаточно различительна. Выберите более подходящие учебные документы; обдумайте возможность удаления этой категории.

– Документы из этой категории попадают в одну или две категории:
  - Скопируйте эти документы в другие категории.
  - Проверьте эту категорию. Есть ли смысл сохранять эту категорию, или категории надо объединить?

• Много **нераспределенных** документов:

– Сравните нераспределенные документы с правильными по:
  - Размер

    Например, если нераспределенные документы короче, можно объединить два таких документа в один более длинный, соответствующий по длине правильным документам.
  - Стилю

    Если стили разные, удалите этот документ.
  - Теме

    Если тема несколько отличается, но все же связана с темой правильных документов, попробуйте найти дополнительные учебные документы по этой теме и выгрузите их в эту или в новую категорию.

  Если получены хорошие значения выхода и точности, но одни и те же документы остаются нераспределенными в последующих итерациях оценки, удалите эти документы.

Процесс оценки нужно запускать повторно, только если таксономия бала изменена, например, если удалена категория, объединены две категории, или было перемещено много документов. Небольшие изменения на уровне категории, такие как добавление новых учебных документов или копирование документов, можно сделать для некоторых категорий без повторения процесса оценки. Учтите, что здесь нет функции отката и вы не сможете отменить внесенные изменения.

Остановите оценку, когда больше нет интересующих вас красных и синих значений, или когда достигнут уровня точности и выхода более 90%.

Обучение каталога

Когда вы оценили таксономию и удовлетворены результатами, нужно обучить ее, используя все учебные документы, чтобы получить особый тип метаданных, называемый моделью категоризации, который можно затем использовать для категоризации новых документов с помощью службы категоризации исследования информации.

Чтобы начать фазу обучения, выберите каталог, для которого нужно провести обучение, щелкните по вкладке **Обучение** и затем нажмите кнопку **Начать обучение**.

В верхней части правой панели выводится имя выбранного каталога. Вы можете увидеть также состояние процесса обучения каталога. Типы состояния обучения:

• Нельзя выполнять категоризацию документов. Выполните обучение каталога, чтобы обновить категоризатор. (Это состояние по умолчанию для вновь созданного каталога. В этом случае дата **Последнее обучение** пуста.) Это новый каталог, который еще не был обучен, или же результаты последнего обучения непригодны, например, после последнего обучения каталога была переименована или удалена какая-либо категория. Попытка использовать службу категоризации для категоризации новых документов приведет к ошибке.

- Изменен набор учебных документов в каталоге. Документы все еще категоризованы в соответствии с последними результатами обучения. Выполните обучение для каталога, чтобы внести изменения в классификатор.
- Идет обучение каталога.
- Каталог в рабочем состоянии, обучение не требуется.

Чтобы остановить процесс обучения, нажмите кнопку **Остановить обучение**. Учтите, что если обучение остановлено, процесс обучения каталога нужно выполнить заново.

Во время обучения каталога в него нельзя выгружать новые учебные документы.

Настройка производительности

При создании записей в каталоге текстовые индексы для атрибутов с возможностью текстового поиска (например, IKF_CONTENT) становятся очень большими и производительность поиска значительно снижается. Для оптимизации хранения и повышения производительности нужно регулярно выполнять реорганизацию текстовых индексов, особенно после значительного изменения индексов.

Реорганизацию текстовых индексов лучше всего запускать в периоды малой загрузки компьютера, например, ночью. Для выполнения организации индексов запустите программу:

- Для Windows: `...\ikf\IkfReorg.cmd`
- Для AIX: `.../ikf/bin/IkfReorg.sh`
- Для Solaris: `.../ikf/bin/IkfReorg`

Параметры программы: IkfReorg <ID_пользователя><Пароль><Имя_БД>

Использование IBM Web Crawler

В этом разделе объясняется, как сконфигурировать IBM Web Crawler. Он устанавливается программой установки EIP, если включен переключатель Возможности.

EIP Версии 8.2 содержит искатель для Web, искатель для Lotus Notes, программу создания сводки для данных, извлеченных из обработанных искателем файлов, документацию в формате HTML, примеры конфигурации и утилиты поддержки. Для IBM Web Crawler (обозначаемого также GCS) требуется Java Версии 1.3 или новее.

IBM Web Crawler - это программа на языке Java, обходящая и исследующая содержимое. Если указать ему содержимое, он получит и исследует это содержимое.

IBM Web Crawler может работать с содержимым в сети предприятия, в объединении таких сетей, в сети Интернет, в базах данных Lotus Notes (напрямую или через Domino) и в локальных файловых системах. В IBM Web Crawler можно легко добавлять новые протоколы. Содержимое может быть любого типа, например, HTML, вложенные файлы Notes и мультимедиа.

IBM Web Crawler может исследовать метаданные и текстовые данные для различных типов содержимого. Например, содержимое HTML может анализироваться по:

- URL
- Заголовок
- Телу документа
- Времени последнего изменения

- Метатегам, таким как автор, ключевые слова, описание и т.д.

Пользователь выбирает нужный вариант анализа из заранее предопределенных для данного типа содержимого. Содержимое и/или анализируемые метаданные сохраняются на локальном диске. IBM Web Crawler может использовать технологию Network Solutions Outside In для извлечения текстовых данных из более 200 типов содержимого - идеальное взаимодействие для прикладных программ поиска. В IBM Web Crawler можно также легко добавить новые программы исследования.

IBM Web Crawler доступен для операционных систем Windows NT 4.0 и Windows 2000. Чтобы установить, сконфигурировать и начать использовать IBM Web Crawler, вам потребуется примерно полчаса. На компьютере PC с тактовой частотой 500 МГц он за секунду получает и анализирует содержимое примерно десяти файлов. Он протестирован для объемов данных до 1 миллиона объектов (200000 записей Lotus Notes). Он поддерживает множество пользователей и множество конфигураций обхода/исследования для каждого пользователя, и позволяет использовать поддержку национального языка, предпочитаемого пользователем.

Возможности IBM Web Crawler

Программа установки устанавливает два файла:

x:/<каталог установки>/run
: Пакетные файлы и примеры конфигурации IBM Web Crawler for the Web.

x:/<каталог установки>/notes-run
: Пакетные файлы и примеры конфигурации Web Crawler for Notes.

x:/<каталог установки>/lib
: Файлы `.jar` и `.zip` IBM Web Crawler и файлы фильтров.

Конфигурирование и запуск IBM Web Crawler для Web

В этом разделе описывается, как сконфигурировать и запустить IBM Web Crawler для Web. IBM Web Crawler для Web получает доступ к серверам HTTP, FTP, новостей или файл-серверам и создает сводки для документов HTML и других объектов. Сводки - это файлы (по одному на документ или объект), содержащие метаданные и полный текст.

Базовая конфигурация

Этот раздел содержит инструкции по редактированию файла конфигурации IBM Web Crawler в формате XML. К системе прилагается два примера файлов конфигурации, которые помогут вам начать работать:

- Файл `config-db2.xml` - для использования IBM Web Crawler с DB2 UDB.
- Файл `config-sample.xml` - для использования IBM Web Crawler без DB2 UDB.

1. Откройте окно командной строки.
2. Перейдите в подкаталог `run` каталога установки IBM Web Crawler. Например, если IBM Web Crawler установлен на сервере Windows, введите `cd x:<cmbroot>\gcs\run`. Если IBM Web Crawler установлен на AIX, введите `cd /usr/lpp/cmb/gcs`.

   Совет: Не забудьте сохранить копию исходного файла. Ошибка в этом файле может повредить IBM Web Crawler. Будьте внимательны при редактировании.
3. Для запуска IBM Web Crawler с базой данных DB2 UDB (этот вариант проще масштабировать, но работает он медленнее) отредактируйте файл `config-db2.xml`. Например, введите в командной строке `edit config-db2.xml`.

4. Для запуска IBM Web Crawler без базы данных DB2 UDB (этот вариант сложнее масштабировать, но работает он быстрее) отредактируйте файл `config-sample.xml`. Например, введите в командной строке `edit config-sample.xml`.

   Чтобы процесс искателя отслеживал *n* адресов URL без базы данных, на компьютере необходимо примерно *n* Кбайт оперативной памяти для хранения найденных метаданных URL. Например, для отслеживания 500000 адресов URL необходимо 512 Мбайт оперативной памяти. Чтобы использовать эту память, отредактируйте файл `crawlweb.bat`, увеличив объем `JVMXmx`.

Конфигурирование соединителя Web Crawler для исследования информации

Чтобы использовать Web Crawler с исследованием информации, то есть чтобы иметь возможность применять функции исследования информации для найденных документов, требуются следующие особые параметры конфигурации, отличающиеся от параметров, заданных в указанных выше примерах файлов конфигурации (`config-sample.xml` и `config-db2.xml`):

```
<globals ...
     max-urls="5000"
     temp-filepool-class="FullPathFilePool"
     summaries-dir="webspaces/ikf/disks/1/"
     summaries-filepool-class="DistributedEIPFilePool"
     ... >
  ...
</globals>

<group-list>
    <group ...>
       ...
         <summarizer-config>
            ...
               <resource-handler content-type="*"
                              summarizable="EipHtmlSummarizable"
                              summary-maker="EipHtmlRawSummaryMaker" />
               ...
         </summarizer-config>
       ...
     </group>
</group-list>
```

Эти параметры входят в пример файла конфигурации для компонента исследования информации, `im-crawler-config-sample.xml`, расположенного в следующем каталоге:

В Windows:

```
<CMBROOT>\samples\java\beans\infomining\webcrawler\
```

В UNIX (AIX и Solaris):

```
<CMBROOT>/samples/java/beans/infomining/webcrawler/
```

Конфигурирование опции DB2 IBM Web Crawler

Чтобы сконфигурировать опцию DB2, нужно создать базу данных. Для этого нужны полномочия администратора DB2. Возможно, потребуется перейти на учетную запись администратора DB2. Для базы данных можно использовать любое имя, допустимое в DB2, но если имя базы данных не `gcs`, потребуется изменить параметр `dbname` в файле конфигурации Web Crawler.

Если у вас есть права администратора баз данных, для создания базы данных можно ввести в командной строке DB2 команду:

db -createdb <*пользователь*><*пароль*>[*имя_базы_данных*]

Если не указать имя_базы_данных, используется имя gcs. После создания базы данных добавьте таблицы IBM Web Crawler, введя команду:

db -createtables<*пользователь*><*пароль*>[*имя_базы_данных*]

Создание базы данных и таблиц IBM Web Crawler необходимо для использования DB2 с IBM Web Crawler.

Чтобы можно было использовать новую базу данных, в разделе urlpool-config файла конфигурации надо задать следующие параметры: dbname:

- Имя созданной выше базы данных, например: gcs.
- User name: Ваше имя пользователя, например: db2admin.
- Password: Пароль этого пользователя, например: db2admin.

Задайте значения свойств database, username и password. Не изменяйте размер кэша и драйвер. Затем в файле нужно задать область для процесса искателя вашей системы.

Задание области для процесса искателя

Ниже приведены параметры файла конфигурации, задающие область для процесса искателя, независимо от того, используете вы DB2 или нет.

Посмотрите в разделе crawler-config следующие параметры и задайте для них нужные вам значения.

seed list
: Один или несколько стартовых полных адресов URL. Эти URL должны быть доступны. Проверьте их при помощи своего браузера, например: http://www.<мой_сайт>.com/

content-type-pattern-list
: Проверять адрес URL, найденный на страницах, только если расширение файла совпадет с каким-либо из шаблонов в списке, например: htm*

include-pattern-list
: Проверять адрес URL, найденный на страницах, только если его страница совпадает с шаблоном в этом списке, например: <мой_сайт>.com

Кроме того, можно задать следующие записи:

recursion-depth
: Максимальная длина цепочки ссылок для проверки из любой начальной точки. Значение -1 означает неограниченную глубину.

exclude-pattern-list
: Отслеживать адрес URL, найденный на страницах, только если его страница не совпадает с шаблоном в этом списке, например: *cgi-bin*

system properties
: Для работы через брандмауэр с компьютера без socks надо также задать в этом файле значения параметра socksProxy.

Запуск IBM Web Crawler

Отредактировав файл конфигурации .xml, сохраните его.

IBM Web Crawler запускается при помощи командного файла crawlweb и файла конфигурации. Откройте окно командной строки и введите:

- Для Windows: crawlweb.bat <ФАЙЛ_КОНФИГУРАЦИИ>

- Для AIX: `crawlweb.sh <ФАЙЛ_КОНФИГУРАЦИИ>`

Для запуска с DB2 UDB введите: `crawlweb config-db2.xml` и нажмите клавишу Enter. Для запуска без DB2 UDB введите: `crawlweb config-sample.xml` и нажмите клавишу Enter.

Совет: Запланируйте регулярные сообщения о ходе выполнения процесса искателя и создания сводок. После проверки адресов назначения сводки записывается в положение, заданное параметром summaries-dir. По умолчанию в сводки записывается исходный объект плюс пролог с метаданными в виде дерева файлов .html. Во время процесса искателя или после него вы можете посмотреть в файле журнала дополнительную информацию.

Расширенная конфигурация

Теперь можно изучить параметры конфигурации. Пример конфигурации смотрите в файле `config-sample2.xml` в разделе Глава 7, “Файлы примеров IBM Web Crawler”, на стр. 105. В этом примере показана конфигурация:

- Потоков искателя и потоков создания сводок
- Графического интерфейса
- Опций записи в журнал
- SOCKS
- Процесса искателя для Lotus Domino
- Нескольких типов содержимого
- Дополнительных исключений
- Использования `InsoSummarizable` для получения сводок по таким объектам, как файлы `.pdf`

Формальное определение возможных параметров в файле конфигурации смотрите в файле `config.dtd`. **Рекомендация:** *Не* редактируйте этот файл. Сделайте копию этого файла и переименуйте ее.

Файл конфигурации IBM Web Crawler

Файл конфигурации - это файл XML, указывающий программе IBM Web Crawler, какие нужно собирать ресурсы Web и как создавать для них сводки. В этом разделе описываются все элементы и атрибуты, которые можно задать в файле `config.xml`. Информацию об использовании IBM Web Crawler for Notes смотрите в разделе “IBM Web Crawler for Notes” на стр. 88.

IBM Web Crawler проверяет, соответствует ли содержимое файла конфигурации требованиям, заданным в файле `gcs-config.dtd`. Если есть серьезные ошибки, например, не заданы URL для работы искателя, IBM Web Crawler завершает работу и выводит сообщение об ошибке. При незначительных ошибках (неизвестный атрибут или значение) программа запишет в журнал предупреждение и продолжит работу. **Рекомендация:** Перед редактированием файла конфигурации сделайте для него резервную копию. Ошибка в этом файле может повредить IBM Web Crawler.

С IBM Web Crawler поставляются примеры файлов конфигурации.

<gcs-config>

Файл `gcs-config` содержит два раздела: **globals** и **group-list**. Пример файла `gcs-config` смотрите в разделе Глава 7, “Файлы примеров IBM Web Crawler”, на стр. 105.

globals Элементы раздела globals задают параметры IBM Web Crawler, например, файловую систему, производительность и сетевую информацию.

group-list

Элементы раздела group-list задают конфигурации процессов искателя и создания сводок для групп, где группа - это набор ресурсов, таких как домены предприятия или сети.

<globals>

Элементы раздела globals задают глобальные параметры IBM Web Crawler. Параметры задаются в виде глобальных атрибутов и дочерних элементов.

В следующем списке определены глобальные атрибуты. Определения глобальных дочерних элементов смотрите в разделе “<logger-config>” на стр. 79.

max-urls

Максимальное число URL, обрабатываемых искателем. Это должно быть положительное целое число; значение по умолчанию - 100000.

summaries-dir

Каталог, в который записываются сводки ресурсов. По умолчанию используется каталог `summaries/`.

summaries-filepool-class

Тип файлового пула, используемого для сводок ресурсов. Он определяет, как создаются имена файлов сводок и какая структура подкаталогов используется (если используется). По умолчанию используется тип `FullPathFilePool` - каталог создается на хосте и затем используются те же структура подкаталогов и имя файла, что и в URL.

num-crawlers

Число используемых потоков искателя. Это должно быть положительное целое число; значение по умолчанию - 20.

num-summarizers

Число используемых потоков сводок. Это должно быть положительное целое число; значение по умолчанию - 5. Для конфигурирования значений num-crawlers и num-summarizers используйте следующие действия:

1. Для числа потоков искателя задайте 1/20 от тактовой частоты компьютера в мегагерцах. Например, в системе с частотой 600 МГц задайте *30*.
2. В качестве числа потоков сводок задайте 1/4 от числа, заданного на шаге 1, например, *8*.
3. Сделайте пробный запуск и посмотрите панель Производительность менеджера задач Windows. Если загрузка CPU *хоть раз* более, чем на секунду, достигнет 100%, вернитесь к шагу 1 и задайте меньшее значение, например, 3/4 от текущего, и делайте так до тех пор, пока загрузка CPU не будет все время меньше 100%.

   Если во время пробного запуска вы видите в информации о состоянии (смотрите параметр text-monitor) постоянно появляющееся сообщение ″Summarizer: ToDo numbers well below the number of configured summarizers″ (Компонент создания сводок: Число создаваемых сводок много меньше числа конфигурированных сводок), можно уменьшать число потоков сводки (чем меньше, тем лучше) и увеличивать число потоков искателя (чем больше, тем лучше), пока не нарушается условие шага 3. Для лучшей производительности используйте насколько возможно более быструю сеть и по возможности разместите сводки, базу данных, временное пространство и журналы на разных дисках.

text-monitor

Если этот элемент имеет значение on, информация о состоянии IBM Web

Crawler выводится каждые пять секунд в `стандартный поток вывода`. Если задано десятичное значение, оно задает интервал (в секундах) между обновлениями этого текстового вывода. Значение по умолчанию - off.

graph-monitor
Если для этого элемента задано значение ″on″, информация о состоянии IBM Web Crawler выводится в графическом пользовательском интерфейсе. Если задано десятичное значение, оно задает интервал (в секундах) между обновлениями информации в графическом интерфейсе. Значение по умолчанию - off.

log-file Задает используемый основной файл журнала. Значение по умолчанию - `log/log.txt`.

Совет: В элементе logger-config можно задать дополнительную информацию о ведении журналов.

log-priority
Задает приоритет журнала по умолчанию. Введите одно из следующих значений: info, warn или error. Значение по умолчанию - warn.

Совет: В элементе logger-config можно задать дополнительную информацию о ведении журналов.

temp-dir
Каталог для временных файлов. **Совет:** Все файлы в этом каталоге могут быть удалены программой IBM Web Crawler. Лучше оставить значение по умолчанию - *x*`:/temp/gcs`.

temp-filepool-class
Тип файлового пула, используемого для временных файлов. **Рекомендация:** Оставьте значение по умолчанию - `TempFilePool`.

content-dir
Каталог, в который IBM Web Crawler записывает файлы содержимого. Обычно значение `content-dir` совпадает со значением `temp-dir`.

content-filepool-class
Тип файлового пула, используемого для файлов содержимого. Обычно совпадает со значением `temp-filepool-class`.

how-often-to-gc
Число URL, обрабатываемых искателем между вызовами сборки мусора. **Рекомендация:** Задайте целое число >= 50. Значение по умолчанию - 100.

max-resource-pool-size
Максимальный размер очереди ресурсов, ожидающих создания сводки. **Рекомендация:** Задайте целое число >= 10. Значение по умолчанию разрешает 10 ожидающих ресурсов на поток сводки.

connect-timeout
Определяет срок ожидания (в миллисекундах) для операции соединения в сети. Значение по умолчанию - 4000. Допустимы значения от 1000 до 60000.

read-timeout
Определяет срок ожидания (в миллисекундах) для операции чтения в сети. Значение по умолчанию - 6000. Допустимы значения от 1000 до 60000.

cookies
Определяет, нужно ли искать в заголовке HTTP опознавательные коды (cookies) и сохранять их в базе данных. Значение по умолчанию - off. Можно разрешить коды опознавания, задав значение on.

locale Определяет язык для сводок и записей журнала. Значение по умолчанию - en_US.

В глобальные дочерние элементы входят logger-configs, urlpool-config и system-properties.

<logger-config>

Файл `logger-config` позволяет управлять тем, какая информация записывается в журнал, форматом журнала и его положением. Файл журнала по умолчанию (`log-file`) и приоритета журнала (`log-priority`) задаются в атрибутах элемента globals. Дополнительную информацию о ведении журналов смотрите в разделе “Ведение журналов в IBM Web Crawler” на стр. 85.

category

Категория ведения журнала, для которой задается конфигурация, например, `gcs.crawler`. Если категория не задана, конфигурируется категория по умолчанию. Учтите, что эти параметры, заданные для определенной категории, будут влиять на все дочерние категории.

priority

Минимальный приоритет сообщений, записываемых журнал. Если не задан, используется приоритет родительской категории (а на верхнем уровне - приоритет по умолчанию, заданный элементом log-priority раздела globals).

log-file

Определяет, куда записывать файл журнала. Если значение начинается с символа ’+’, этот файл журнала будет использоваться как дополнительный вместе с другими (родительскими) файлами журналов. Если не задан, будет использоваться файл журнала родительской категории (а на верхнем уровне - файл журнала по умолчанию, заданный элементом log-file раздела globals).

Совет: Будьте внимательны и не задавайте один файл журнала для нескольких категорий ведения журнала, поскольку в таком случае файлы будут перезаписываться.

log-layout

Определяет макет, используемый для вывода сообщений в файл журнала.

<urlpool-config>

В файле `urlpool-config` задается конфигурация компонента программы IBM Web Crawler, в котором сохраняются URL. Есть несколько вариантов пула URL. Он может храниться в памяти, можно использовать DB2 или же особую версию с небольшой памятью, сохраняющую меньше информации для каждого URL. Если элемент `urlpool-config` не задан, пул URL хранится в памяти. У элемента `urlpool-config` могут быть дочерние элементы `urlpool-param` в которых, например, может задаваться информация о базе данных.

urlcontainer-class

Тип используемого контейнера URL. Задайте:

- DB2URLContainer, чтобы использовать DB2 UDB
- MemoryURLContainer, чтобы не использовать DB2 UDB (значение по умолчанию).
- BigMemoryURLContainer, чтобы не использовать DB2 UDB и использовать дополнительную память (для хранения некоторых ссылочных URL и другой информации).

urlcollection-class

Тип используемого собрания URL. Задайте:

- DB2URLCollection, чтобы использовать DB2 UDB

- MemoryURLCollection, чтобы не использовать DB2 UDB (значение по умолчанию).
- BigMemoryURLCollection, чтобы не использовать DB2 UDB и использовать дополнительную память (для хранения некоторых ссылочных URL и другой информации).

<urlpool-param>

Используется для передачи параметров собранию, заданному `urlcollection-class`. Например, смотрите информацию соединения с базой данных в примере конфигурации для использования DB2 UDB в разделе Глава 7, “Файлы примеров IBM Web Crawler”, на стр. 105.

name Определяет имя параметра.

value Определяет значение параметра.

Совет: Будьте осторожны при использовании этих параметров - они не проверяются на наличие ошибок.

<system-properties>

В system-properties задается список параметров, определяющих свойства системы.

<property>

Свойство; пример смотрите в конфигурации для использования шлюза SOCKS в примере расширенной конфигурации.

name Имя параметра.

value Значение параметра.

Можно также сконфигурировать IBM Web Crawler для доступа к внешним серверам через шлюз PROXY:

```
<system-properties>
  <property name="proxySet" value="true"/>
  <property name="proxyHost" value="proxy.имя_хоста"/>
  <proxy port name="proxyPort" value="80"/>
</system-properties>
```

Совет: Значения этих параметров не проверяются на ошибки, поэтому будьте осторожны при их использовании.

<group-list>

Элемент group-list - это список из одного или нескольких элементов group.

<group>

Элемент group представляет одну группу ресурсов, для которых используется собственная конфигурация искателя и создания сводок. Каждая группа должна иметь уникальный атрибут имени и как минимум один дочерний элемент crawler-config, задающий объекты для обработки. Если вы не хотите использовать для этой группы создание сводок по умолчанию, для нее можно задать дочерний элемент summarizer-config. **Совет:** Перекрытие групп (две или несколько групп, содержащих один и тот же URL) может дать непредсказуемые результаты. URL, входящие в несколько групп, связываются только с первой группой, в которой они найдены.

name - уникальное имя этой группы (обязательный параметр).

<crawler-config>

Эти правила используются для задания области работы искателя. Искатель берет каждый URL из списка seed-list, выполняет синтаксический разбор URL и добавляет в список URL для обработки те URL, которые:

- удовлетворяют по крайней мере одному правилу в списке `content-type-pattern-list`
- и удовлетворяют по крайней мере одному правилу в списке `include-pattern-list`
- и не удовлетворяют ни одному из правил в списке `exclude-pattern-list`.

Для элемента crawler-config также требуется один атрибут: **recursion-depth**. Этот атрибут определяет число ссылок от начального URL, для которых выполняется обработка. Значение по умолчанию -1, что означает неограниченную глубину обработки.

<seed-list>
Список начальных URL (возможно, с информацией аутентификации).

<seed>
Начальный URL для работы искателя, с атрибутом URL и, возможно, информацией аутентификации. Каждый такой URL должен быть абсолютным, например, `http://<your.server>.com/.` Не используйте URL с переадресацией, недоступные или указывающие на нетекстовые страницы. Удобно задавать здесь URL страницы, на которой вы разместили нужные начальные URL. Такую страницу легко изменять, просматривать и проверять с помощью браузера.

URL Начальный URL для работы искателя.

<authentication>
Дополнительная информация аутентификации, посылаемая начальному URL, защищенному с помощью *Basic Authentication* (как определено в rfc2617).

username
Имя пользователя, используемое для аутентификации.

password
Пароль, используемый для аутентификации.

Например:

```
<seed url="http://your.server.com/"><authentication username="me"
password="mine"/></seed>
```

<content-type-pattern-list>
Это список шаблонов, определяющих типы обрабатываемого содержимого по расширениям файлов. Каждый файл с расширением (.html, .gif, .doc и т.п.), соответствующим какому-либо из шаблонов в этом списке, считается удовлетворяющим условию. Файлы без расширений считаются по умолчанию удовлетворяющими условию. Если список `content-type-pattern-list` не задан или задан пустой список, принимаются только URL без расширения файла.

<include-pattern-list>
Это список шаблонов, определяющих обрабатываемые URL, например, по имени сервера или домена. Каждый URL, соответствующий какому-либо из шаблонов `url-obj-pattern`, `url-regex-pattern`, `url-name-pattern` или `url-predicate-pattern` в это списке, считается удовлетворяющим условию. Если список `include-pattern-list` не задан или задан пустой список, принимаются все URL.

<exclude-pattern-list>
Это список шаблонов, определяющих URL, которые не должны обрабатываться. Если URL соответствует какому-либо из шаблонов `url-obj-pattern`, `url-regex-pattern`, `url-name-pattern` или `url-predicate-pattern` в это списке, обработка для него не будет выполняться. Если список `exclude-pattern-list` не задан или задан пустой список, никакие URL не будут отвергаться.

<url-obj-pattern>

Это шаблон, в котором задаются различные части URL (протокол, хост и т.д.) с символами подстановки. Он может использоваться в exclude-pattern-list и в include-pattern-list. В начале и/или в конце шаблона для каждой части URL может быть символ подстановки `'*'`, соответствующий любому значению. Однако символы подстановки нельзя использовать в середине шаблонов. При сравнении значений регистр символов не учитывается. Все пропущенные шаблоны частей URL автоматически считаются совпадающими.

Ниже показан пример того, как Java и IBM Web Crawler делят на части URL `http://www.ibm.com/products/index.html?query#ref`:

- Протокол: `http`
- Хост: `www.ibm.com`
- Порт: `-1` (не задан)
- Файл: `/products/index.html?query`
- Путь: `/products/index.html`
- Каталог: `/products/`
- Имя файла: `index.html`
- Расширение: `.html`
- Запрос: `query`
- Ссылка: `ref`

В следующем списке более подробно описаны все элементы `url-obj-pattern`:

протокол
Шаблон подстановки, которому должен соответствовать протокол URL, например, `http`.

хост Шаблон подстановки, которому должен соответствовать хост URL, например, `*.ibm.com`.

порт (port)
Шаблон подстановки, которому должен соответствовать порт URL, например, `80`.

файл Шаблон подстановки, которому должен соответствовать файл URL, например, `*.htm*.`. Часть URL, соответствующая файлу, начинается с первого слэша после имени хоста и может включать запрос, но не ссылку. Часть *файл* в `http://www.ibm.com/products/index.html?query#ref` - это `/products/index.html?query`.

путь Шаблон подстановки, которому должен соответствовать путь URL, например, `*.html`. Часть URL, соответствующая пути, начинается с первого слэша после имени хоста и не включает запрос и ссылку. В нашем примере часть *путь* в `http://www.ibm.com/products/index.html?query#ref` - это `/products/index.html`

каталог
Шаблон подстановки, которому должен соответствовать каталог в URL, например, `/products/`. Каталог - это часть пути от первого до последнего слэша. В нашем примере часть *каталог* в `http://www.ibm.com/products/index.html?query#ref` - это `/products/`. В эту часть не входят запрос или ссылка. Имейте в виду, что URL, в которых нет последнего слэша, например, `http://www.ibm.com/products`, не будут правильно соответствовать значению `каталога` products. В этом примере с неправильным URL значением каталога будет `/`.

имя файла

Шаблон подстановки, которому должно соответствовать имя файла в URL, например, `index.html`. Имя файла - это часть пути после последнего слэша. В нашем примере часть *имя файла* в `http://www.ibm.com/products/index.html?query#ref` - это `index.html`. В имя файла не входят запрос и ссылка.

расширение

Шаблон подстановки, которому должно соответствовать расширение файла в URL, например, `htm*`. Рекомендуется по возможности лучше использовать `content-type-pattern-list`.

запрос Шаблон подстановки, которому должен соответствовать запрос в URL.

ссылка

Шаблон подстановки, которому должна соответствовать ссылка в URL (не используется для HTTP). Например, `<url-obj-pattern host="*.ibm.com"/>` будет соответствовать страницам HTML на любом сайте IBM.

<url-regex-pattern>

В шаблоне `url-regex-pattern` URL задается регулярным выражением. Этот шаблон может использоваться в `exclude-pattern-list` и в `include-pattern-list`. Он использует пакет `com.ibm.regex (regex4j)` и поддерживает большую часть возможностей регулярных выражений Perl 5. В этом шаблоне можно задать два регулярных выражения: одно выражение, которому URL *должен* соответствовать, и второе, которому URL *не должен* соответствовать. Могут быть заданы и другие опции, например, `i` для сравнения без учета регистра символов. Подробную информацию смотрите в описании регулярных выражений `Regex4j`.

match Регулярное выражение в формате Perl 5, которому URL должен соответствовать.

no-match

Регулярное выражение в формате Perl 5, которому URL должен не соответствовать.

Опции Необязательные модификаторы, например, `i` для сравнения без учета регистра символов.

Например, шаблону `<url-regex-pattern match="^http://www\.ibm\.com/.*\.html?$"/>` будут соответствовать страницы HTML на главном сайте IBM.

<url-name-pattern>

Это простой шаблон с символами подстановки, задающий весь URL или расширение файла URL. Он может использоваться в `content-type-pattern-list`, `include-pattern-list` и `exclude-pattern-list`. В начале и/или в конце строки шаблона может быть символ подстановки `'*'`, соответствующий любому значению. Однако символы подстановки нельзя использовать в середине имени. При сравнении значений регистр символов не учитывается.

Например, шаблону `<url-name-pattern name="*.ibm.com/*"/>` будут соответствовать все файлы на сайте IBM, но шаблон `<url-name-pattern name="*.ibm.com/*.html"/>` недопустим, поскольку в нем символ подстановки стоит в середине строки.

name задает шаблон с символами подстановки, которому должна соответствовать строка URL; в начале и/или конце шаблона могут быть необязательные символы подстановки `'*'`.

<url-predicate-pattern>

Этот шаблон загружает класс Java UnaryPredicate, используемый для проверки соответствия URL. Он может использоваться в exclude-pattern-list и в include-pattern-list. В этом классе должен быть метод `public boolean execute(URL url)`, возвращающий значение true, если URL соответствует предикату.

class задает полное имя класса `UnaryPredicate`.

<summarizer-config>

Это конфигурация компонента создания сводок со списком дочерних элементов resource-handler. В настоящее время для каждой группы может быть только один элемент `summarizer-config`.

<resource-handler>

Определяет тип сводок, создаваемых для ресурсов (таких, как Web-страницы или статьи групп новостей) в зависимости от типа содержимого ресурса, например, `(text/html)` или расширения имени файла (`htm`). Если ресурс готов к созданию сводки, IBM Web Crawler по очереди проверяет значения элементов resource-handler и использует первое значение, соответствующее типу содержимого или расширению файла. Если соответствие не найдено, по умолчанию используются `Copy2RdfSummarizable` и `Copy2RdfSummaryMaker`. Эти значения по умолчанию можно переопределить, добавив в начало списка элемент resource-handler без шаблонов content-type и file-extension.

У элемента resource-handler может также быть дочерний элемент `summarizer-param`, задающий специальные параметры для передачи классу `SummaryMaker`.

content-type
: Шаблон с символами подстановки, которому должен соответствовать тип содержимого ресурса, например: `*htm*`

file-extension
: Шаблон с символами подстановки, которому должно соответствовать расширение файла ресурса, например: `htm*`

summarizable
: Имя класса Summarizable для ресурса, например: `HtmlRawSummarizable`

summary-maker
: Имя класса SummaryMaker для ресурса, например: `HtmlRawSummaryMaker`

В шаблонах `content-type` и `file-extension` можно использовать символы подстановки. В начале и/или в конце строки шаблона может быть символ подстановки `*`, соответствующий любому значению. Однако символы подстановки нельзя использовать в середине шаблона. При сравнении значений регистр символов не учитывается.

Условие элемента `resource-handler` удовлетворяется, если выполняется соответствие для обоих шаблонов `content-type` и `file-extension` (если шаблон не задан, соответствие для него выполняется всегда), поэтому условию

```
<resource-handler content-type="*htm*"
summarizable="*HtmlRawSummarizable" summary-maker=
"HtmlRawSummaryMaker"/>
```

будут удовлетворять все файлы с типом содержимого `text/html`, независимо от расширения файла.

Для элементов summarizable и summary-maker можно не задавать полный путь классов, если эти классы входят в пакет com.ibm.IBM Web Crawler.summarizer.resource.

<summarizer-param>

Специальные параметры, передаваемые классу SummaryMaker. Они зависят от конкретного класса.

name Имя параметра.

value Значение параметра.

Совет: Значения этих параметров не проверяются на ошибки, поэтому будьте осторожны при их использовании.

Ведение журналов в IBM Web Crawler

Это введение в средства ведения журналов IBM Web Crawler.

IBM Web Crawler имеет мощные средства управления записываемой в журналы информацией, положением журналов и их форматом. Например, можно записывать коды ответов для каждой обработанной искателем страницы в один файл, информацию о состоянии IBM Web Crawler (число обработанных URL, число работающих потоков и т.д.) - в другой файл, URL, для которых созданы сводки - в третий файл, все предупреждения IBM Web Crawler - в четвертый файл, а все сообщения регистрации в пакете сетевых утилит - в еще один файл, чтобы использовать его для отладки.

Пример файла анализа регистрации смотрите в разделе “Пример анализа файла журнала IBM Web Crawler” на стр. 107.

Использование журналов

Регистрация в журналах позволяет сохранять информацию о сети, Web, работе искателя и создании сводок, передавать эту информацию другим программным компонентам и использовать ее для отладки IBM Web Crawler.

Данные учета, сгенерированные процессами искателя и исследования информации, позволяют узнать много интересного, например, обнаружить неправильно сконфигурированные сервера, отсутствующие страницы или узнать число объектов с конкретным типом содержимого. Сценарий Perl loganalysis.pl - это пример записи в журнал сводных данных учета. Прикладным программам может требоваться информация от IBM Web Crawler, например, о моменте удаления содержимого.

Конфигурирование регистраторов

В файле конфигурации IBM Web Crawler можно задать конфигурации одного или нескольких регистраторов для ведения журналов. Атрибуты log-priority и log-file элемента globals задают правила записи в журнал по умолчанию.

Чтобы расширить правила записи в журнал, создайте оператор logger-config в виде дочернего элемента для globals. Каждый оператор выбирает подмножество сообщений журнала IBM Web Crawler, направляет их в отдельный файл и записывает их в определенном формате. Подмножество сообщений журнала выбирается по атрибутам priority (приоритет) и category (категория). Допустимые значения приоритета: trace, debug, info и warn (значения регистронезависимы).

- Значение приоритета определяет объем записываемой информации; приоритет trace задает наибольший объем информации.
- trace и debug - это служебные уровни, сообщения в них записываются только на английском языке.

- `info` и `warn` - пользовательские уровни, на них поддерживаются национальные языки.
- При приоритете `info` генерируется много сообщений. Чтобы уменьшить число сообщений, задайте приоритет `warn`.

Примеры конфигурации записи в журнал

Регистрировать гиперссылки из/в, без информации о дате/времени/потоке, в файл `log/fromto.txt`

```
<logger-config category="gcs.url.fromto" priority="info"
 log-layout="%m\n" log-file="log/fromto.txt"/>
```

Записывать объекты, для которых составлены сводки, в файл `log/resources.txt`

```
logger-config category="gcs.summaries.list.resource"
 priority="info" log-file="log/resources.txt"/
```

Записывать пропущенные URL и причины пропуска

```
<logger-config category="gcs.url.skipped"
priority="info" log-file="log/urls_skipped.txt"/>
```

Записывать специально обработанные коды ответа HTTP

```
<logger-config category="gcs.http.302"
priority="info" log-file="log/urls_redirected.txt"/>
<logger-config category="gcs.http.404"
priority="info" log-file="log/urls_not_found.txt"/>
```

Записывать все сообщения в категории процесса сводки, включая их приоритеты

```
 <logger-config category="gcs.summarizer"
priority="TRACE" log-file="summarizer_trace.txt"
log-layout="%d: %t: %c: %p: %m\n"/>
```

Устранение неисправностей

Если у вас возникли проблемы, прежде всего проверьте:

Доступны ли страницы из списка начальных URL?
Эти страницы должны существовать (избегайте переадресации) и быть доступны с вашей системы (и через SOCKS, если используется SOCKS).

Страницы в списке начальных URL содержат обычный HTML?
В качестве начальных страниц лучше не использовать страницы с фреймами, использующие Flash, Javascript и другие аналогичные элементы. Выбираете простые HTML-страницы.

Если используется DB2 UDB, был ли уже выполнен процесс искателя?
DB2 UDB сохраняет результаты работы искателя. Если все страницы были обработаны, DB2 UDB будет неактивной. Чтобы еще раз запустить искатель, используйте команду `db -emptytables`.

Если используется DB2 UDB, правильно ли заданы параметры доступа к базе данных в файле конфигурации?
Если не удается соединиться с базой данных, работа искателя будет неудачной.

Внимательно ли вы проверили значения в файле конфигурации?
Ошибки могут нарушить работу IBM Web Crawler. Проверили ли вы ограничивающие значения max-urls и recursion-depth?

Ошибка повторяется?
Отредактируйте файл конфигурации, изменив приоритет регистрации на ″debug″. Теперь еще раз запустите Web Crawler и после завершения его работы посмотрите файл журнала.

Выбор генераторов сводок

Генератор сводок берет ресурс (например, Web-страницу) или хост (например, Web-сервер) и создает файл, содержащий интересующую вас информацию в удобном формате.

IBM Web Crawler содержит различные генераторы сводок, которые обрабатывают разные типы содержимого, извлекают из ресурсов разные типы данных и записывают выходные данные в файлы разных форматов. В этом разделе описываются возможности и требования доступных генераторов сводок. Если ни один из них не удовлетворяет вашим потребностям, можно написать свой собственный генератор сводок.

В IBM Web Crawler есть два типа генераторов сводок. Ресурсный генератор создает сводку для одного ресурса, например, Web-страницы, а серверный генератор сводок - для хоста, например, Web-сервера. В настоящее время можно конфигурировать только ресурсные генераторы сводок.

Выбор ресурсного генератора сводок

Что нужно учитывать при выборе или написании ресурсного генератора сводок:

- Какой будет входной формат? (Web-страница HTML, PDF, документ WordPro, файл XML)
- Какие метаданные нужно извлекать? (Заголовок HTTP, название документа, ссылки с надписями, основной текст)
- Какой требуется выходной формат? (XML, HTML, RDF)

Чтобы задать генератор сводок для конкретного типа ресурсов, используйте элемент resource-handler в файле `config` IBM Web Crawler. Сначала задайте тип содержимого и/или расширение файла, для которых будет использоваться этот генератор сводок. Затем задайте классы Java Summarizable и SummaryMaker, используемые для создания сводки. Класс summarizable представляет ресурс, для которого нужно сделать сводку, а класс summary-maker - тип создаваемой сводки.

Генератор сводок по умолчанию (Copy + RDF)

Генератор сводок Copy + RDF применяется для всех объектов, для типа содержимого (content-type) которых явно не задана обработка другим компонентом создания сводок. Этот генератор сводок можно использовать для любого типа ресурсов. Он создает два файла: в первый файл записывается копия исходного ресурса, а во второй файл - сводка RDF, содержащая исходный URL, имя сохраненного файла и информацию заголовка HTTP. Этот генератор сводок можно также задать явно, введя значения `DefaultSummarizable` и `Copy2RdfSummaryMaker`.

Генератор сводок для страниц HTML (Raw HTML)

Для ресурсов HTML генератор сводок Raw HTML просто создает копию исходного файла, в комментарии в начале которого записываются URL и информация заголовка HTTP. Его использование задается значениями `HtmlSummarizable` и `HtmlRawSummaryMaker`.

```
<resource-handler content-type="*htm*"
                  summarizable="HtmlSummarizable"
                  summary-maker="HtmlRawSummaryMaker" />
```

Генераторы сводок для страниц HTML (EIP HTML)

Для ресурсов HTML генератор сводок Raw HTML просто создает копию исходного файла, в комментарии в начале которого записываются URL и информация заголовка HTTP. Его использование задается значениями `EIPHtmlSummarizable` и `EIPHtmlRawSummaryMaker`.

```
<resource-handler content-type="*htm*"
                    summarizable="EIPHtmlSummarizable"
                    summary-maker="EIPHtmlRawSummaryMaker" />
```

Генератор сводок для страниц HTML (No write HTML)

Этот генератор сводок просматривает HTML и последующие ссылки, но не записывает в файл никакой сводки. Это может быть полезно, если, например, нужно обработать все файлы PDF на каком-то сайте (используя генератор сводок ″INSO to XML″), не сохраняя файлы HTML. Его использование задается значениями `HtmlSummarizable` и `NoWriteSummaryMaker`.

```
<resource-handler content-type="*htm*"
                                summarizable="InsoSummarizable"
                                summary-maker="InsoSummaryMaker" />
```

Генератор сводок для других типов содержимого (INSO to XML)

Этот генератор сводок создает сводки XML для более чем 200 типов ресурсов, например, документов Microsoft Word, файлов PDF, презентаций PowerPoint и других. Такая сводка содержит определенные метаданные и текст тела документа, извлеченный фильтрами Network Solutions INSO (требуется лицензия INSO). Его использование задается значениями `InsoSummarizable` и `InsoSummaryMaker`.

```
<resource-handler content-type="pdf"
                                summarizable="InsoSummarizable"
                                summary-maker="InsoSummaryMaker" />
```

Другие генераторы сводок

Если вам нужно генерировать сводки для других типов ресурсов, исследовать другие данные или создавать выходные данные в других форматах, обратитесь в IBM или создайте собственный генератор сводок.

IBM Web Crawler for Notes

В этом разделе рассказывается, как сконфигурировать и запустить IBM Web Crawler for Notes. IBM Web Crawler for Notes обращается к базам данных Notes и создает сводки для документов и вложенных файлов Notes. Сводки - это файлы формата XML (один файл для каждого документа или вложенного файла), содержащие объект и полный текст.

Предварительные требования

Для работы IBM Web Crawler for Notes требуются:

- Lotus Notes Версии 5.0.5 или новее.
- PKZIP Версии 2.50 (если нужно обрабатывать вложенные файлы формата zip).

Тестовый запуск искателя

Выберите **Пуск →Программы →Командная строка**. В открывшемся окне перейдите в подкаталог `notes-run` каталога, где установлен IBM Web Crawler for Notes. Например:

```
cd c:\<каталог установки>\gcs\notes-run
```

Процесс искателя для данных Notes управляется двумя файлами:

- Списком источников, где задаются базы данных Notes, с которыми может работать искатель. Здесь задаются имена серверов Notes, IP-адреса, имена файлов .nsf и т.д. Например, в списке источников вы можете задать 34 базы данных Notes; конкретная база данных, с которой будет работать искатель, задается в файле конфигурации, описанном ниже. Список источников может быть или файлом .xml, или базой данных Notes (файлом .nsf).

- Файлом конфигурации, в котором задаются: список источников, с каким именно источником должен работать искатель, какие типы вложенных файлов он должен обрабатывать, выходные форматы и т.д. Файл конфигурации - это всегда файл .xml.

Чтобы проверить правильность установки IBM Web Crawler, запустите его для тестовой базы данных. Используя текстовый редактор, убедитесь, что в списке источников `testSources.xml` заданы правильные путь и имя файла базы данных `test.nsf`, которая расположена в подкаталоге `notes-run` каталога установки IBM Web Crawler. Сделайте резервные копии исходных файлов. **Рекомендация:** *Будьте внимательны при редактировании*: ошибки в этом файле помешают работе IBM Web Crawler. Сохраните все изменения.

Протестируйте свою установку, запустив искатель для поставляемой базы данных `test.nsf`. Введите: `crawlNotes crawlTestXml`

Файл `crawlNotes.bat` запустит IBM Web Crawler с `crawlTestXml` в качестве файла конфигурации; расширение `.xml` автоматически добавляется в конец имени файла конфигурации. IBM Web Crawler должен сообщить об обработке и создании сводок для двух документов (каждый из них с вложенным файлом).

Когда IBM Web Crawler завершит свою работу, вы можете просмотреть сводки в каталоге сводок и файлы журналов искателя в каталоге журналов, заданном в файле конфигурации.

Конфигурирование процесса искателя для Notes

После успешного выполнения тестового запуска искателя можно использовать искатель для других баз данных.

1. Создайте список баз данных источников. Добавьте в файл источников базы данных Notes, с которыми должен работать искатель.

   Чтобы задать базы данных Notes для искателя в файле XML, откройте в текстовом редакторе файл `testSources.xml`. Чтобы задать базы данных Notes для искателя в базе данных Notes, с помощью Notes откройте и измените базу данных Notes `testSources.nsf`. Параметры, которые можно задать в файлах источников, объясняются в разделе Редактирование списка источников.
2. Задайте конфигурацию искателя. Вам нужно отредактировать файл конфигурации в формате XML.
   - Если список источников задан в файле XML, откройте в редакторе файл `crawlTestXml.xml` и задайте в параметре `sourcesInXmlFile` имя файла источников.
   - Если список источников задан в базе данных Notes, откройте в редакторе файл `crawlTestNsf.xml` и задайте в параметре `sourcesInNotesDB` имя базы данных источников. Параметры, которые можно задать в файлах конфигурации, объясняются в разделе “Файл конфигурации IBM Web Crawler” на стр. 76.

Задав список источников и конфигурацию, запустите IBM Web Crawler: `crawlNotes ваш_файл_конфигурации`

Если же список источников задан в базе данных Notes, запустите IBM Web Crawler, как в следующем рабочем примере: `crawlNotes crawlTestNsf`

Когда IBM Web Crawler for Notes завершит свою работу, можно просмотреть сводки в каталоге сводок и файлы журналов искателя в каталоге журналов, заданном в файле конфигурации.

Параметры списка источников

Список источников содержит описания баз данных Notes, с которыми может работать искатель. Списки источников в формате XML содержат элемент notesDataSources с одним или несколькими элементами oneDBinfo. Каждый элемент oneDBinfo содержит:

id Числовой идентификатор этой базы данных. Он используется в параметре range в файле конфигурации.

serverName
Имя сервера этой базы данных. Для локальной базы данных используйте пустую строку `""`.

pathAndFileName
Полный путь и имя файла базы данных на сервере. В конце имени файла задайте расширение `.nsf`.

viewName
Имя представления Notes для базы данных, с которой будет работать искатель.

ipAddress
Необязательный. IP-адрес сервера; если он задан, DNS не используется. Если DNS не может найти нужное имя сервера, вы можете задать здесь IP-адрес этого сервера. В Windows IP-адрес можно определить с помощью команды `nslookup имя_сервера`.

dateLastCrawled
Необязательный. Дата последней обработки базы данных искателем. Это значение изменяется автоматически (если в файле конфигурации для параметра `updateDateLastCrawled` не задано значение ″no″).

tries Необязательный. Число повторов процесса искателя для базы данных, если процесс искателя не завершен успешно (превышен срок ожидания).

fieldSubstitutions
Отображения, задающие, на что заменяются имена полей базы данных Notes в выходном документе XML. Содержит один или несколько элементов substitute, каждый с двумя атрибутами:

- Original: имя поля, которое будет заменяться в выходном документе XML (если оно существует)
- Replace: новое имя поля, на которое будет заменяться исходное имя поля в выходном документе XML.

Списки источников в базе данных Notes можно проверить и изменить с помощью клиента Notes. Запустите Notes и выберите **File →Database →Open** (Файл →База данных →Открыть). Нажмите кнопку **Browse** (Обзор), чтобы найти и открыть базу данных `testSources.nsf` в каталоге *x*`:\<каталог установки>/gcs/notes-run`. Поставляемую тестовую базу данных также можно проверить и изменить с помощью клиента Notes. Запустите Notes и выберите **File →Database →Open** (Файл →База данных →Открыть). Нажмите кнопку **Browse** (Обзор), чтобы найти и открыть базу данных `test.nsf` в каталоге *x*`:\<каталог установки>/gcs/notes-run`.

Параметры файла конфигурации

Ниже описываются параметры, которые можно задать в файле конфигурации. Параметры, для которых приведены значения по умолчанию, задавать не обязательно.

Список источников задается в элементах sourcesInXml или sourcesInNotesDB. В них задаются списки источников в файлах формата XML или базах данных, соответственно; списки источников содержат информацию о базах данных для работы искателя.

Элемент runInfo содержит параметры, управляющие отдельным запуском. Это значит, что они применяются ко всем базам данных Notes, обрабатываемым в данном запуске искателя:

rangeSpecify
Идентификаторы баз данных для работы искателя. Идентификаторы - это числа, заданные в поле id в списке источников. Можно задать список отдельных идентификаторов и/или диапазонов через запятую, например ″1-4, 15, 25-31″.

SummaryDirectory
Задает корневой каталог для выходных сводок. Сводки записываются в подкаталоги этого каталога.

MaxThreads
Задает число параллельных потоков искателя. Для обработки каждой базы данных Notes используется отдельный поток. Несколько баз данных обрабатываются параллельно.

doIncrementalCrawl
Значение по умолчанию - ″no″. Если задано значение ″yes″, будут обрабатываться только документы Notes, созданные или измененные после даты, заданной параметром 'summarizeThisDateAndLater'. Если параметр 'summarizeThisDateAndLater' не задан, искатель будет использовать для конкретной базы данных значение поля DateLastCrawled, заданное для нее в списке источников. Если doIncrementalCrawl=no, будут обрабатываться все документы, независимо от даты.

summarizeThisDateAndLater
Формат этого поля: ММ/ДД/ГГГГ чч:мм AM/PM пояс, например, 01/01/2000 01:11 PM PDT. Если дата и время не заданы, сводка создается для всех документов, дата которых позже даты последнего запуска искателя, записанной в списке источников (если doIncrementalCrawl имеет значение ″yes″), или без ограничений (если doIncrementalCrawl имеет значение ″no″).

detachAttachments
Значение по умолчанию - ″yes″. Если значение ″yes″, выделять и обрабатывать вложенные файлы. Типы обрабатываемых вложенных файлов перечисляются в файле конфигурации. Если значение ″no″, вложенные файлы игнорируются.

attachmentFilenameFormat
Значение по умолчанию - ″l″ (длинное). Другое возможное значение - ″s″ (короткое). В длинном имени файла кодируются тип, сервер, имя базы данных и id Notes. В коротком имени файла кодируются тип и id Notes.

processAttachmentsAfterwards
Значение по умолчанию - ″no″. Если задано значение ″yes″, при создании сводок для базы данных Notes не будут создаваться сводки для вложенных файлов. Вместо этого в файл notesCrawl-attachments.bat будет помещена запись для каждого вложенного файла, содержащая команду для создания сводки этого файла. Позже вы можете написать и выполнить пакетный файл, создающий сводки для вложенных файлов и затем удаляющий эти файлы. Для такой отложенной обработки вложенных файлов обычно требуется большой объем дискового пространства.

saveAttachmentFiles
Значение по умолчанию - ″no″. Если задано значение ″yes″, исходные вложенные файлы не будут удаляться после обработки. Эта опция используется, только если `processAttachmentsAfterwards` имеет значение ″no″. Если для `processAttachmentsAfterwards` задано значение ″yes″, необходимо дисковое пространство для сохранения вложенных файлов.

MaximumNumberOfDetachingErrors
Значение по умолчанию - 10. Максимальное число ошибок при обработке вложенных файлов (например, нехватка дискового пространства при сохранении вложенных файлов), не вызывающих прекращения работы искателя.

saveURLsToFile
Значение по умолчанию - ″no″. Если задано значение ″yes″, URL, обнаруженные в документах Notes, будут записываться в файл с именем вида:
`имя_базы_данных(без пути и расширения .nsf) + ".html".`

updateDateLastCrawled
Значение по умолчанию - ″yes″. Если значение ″no″, значение параметра dateLastCrawled в файле источников не будет обновляться.

tempDirectory
Значение по умолчанию - `c:\temp`. Этот каталог используется для записи всех временных файлов.

logSummaryDirectory
Значение по умолчанию - log. Задает каталог, в который сохраняются файлы журналов.

loggerPriority
Значение по умолчанию - info. Значение этого параметра определяет приоритет регистратора. Допустимые значения (от высшего к низшему): error, warn или info. Например, если задан приоритет регистратора warn, в журнал будут записываться только сообщения с приоритетами warn и error.

Элемент attachments содержит элементы include, задающие расширения вложенных файлов, которые нужно обрабатывать, например `.prz`.

Исключение сервера из области работы IBM Web Crawler

В целях защиты и повышения производительности администратор EIP может решить исключить из области работы искателя содержимое определенных серверов или страниц. Для этого может потребоваться возможность ограничивать активность искателей на серверах и страницах.

Используя файл *правил доступа*, можно указать искателю IBM Web Crawler, чтобы он не заходил на определенные серверы и страницы. Этот файл строится в соответствии со стандартом *A Standard for Robot Exclusion* (смотрите `http://info.webcrawler.com/mak/projects/robots/norobots.html`).

- IBM Web Crawler запрашивает файл правил доступа http://*ваш_сервер*/robots.txt перед началом работы с сервером и периодически во время работы.
- Этот файл состоит из строк вида
  поле:<необязательный_пробел>*значение*<необязательный_пробел>

  Если *поле* - `User-Agent`, а *значение* - `IBM-WebCrawler` или `*`, последующие строки `Disallow` (до следующей строки `User-Agent`) задают конкретные запрещенные адреса. Это может быть полный или частичный путь; если какой-либо адрес начинается с этого значения, данные с него не будут получены.

Например:

```
Disallow: /help
```

закрывает и `/help.html`, и `/help/index.html`.

```
Disallow: /help/
```

закрывает `/help/index.html`, но разрешает `/help.html`.

Пустое значение разрешает получать данные со всех адресов.

- Строки можно отделять друг от друга пустыми строками.
- Можно вставить свое замечание, введя символ #. Вся остальная часть строки будет рассматриваться как комментарий.

Ниже приводятся несколько примеров:

- Этот файл `/robots.txt` задает, что все роботы должны избегать этого сервера.

```
# запретить всем

User-agent: *
Disallow:   /
```

- Этот файл `/robots.txt` задает, что только IBM Web Crawler может работать с этим сервером и что для него нет ограничений.

```
# разрешить только IBM

User-agent: *
Disallow:   /
User-agent: IBM-WebCrawler
Disallow:   # запретов нет
```

- Этот файл `/robots.txt` задает, что все роботы должны избегать адресов в подкаталогах `temp`, `development` и `testing` каталога htmldocs, что пользовательский агент с именем IBM-WebCrawler может работать с подкаталогами development и testing (они не запрещены для него) и что роботам `xyz` и `wxyz` доступ запрещен полностью.

```
# более реальный пример

User-agent: *
Disallow:   /htmldocs/temp
Disallow:   /htmldocs/development
Disallow:   /htmldocs/testing

User-agent: xyz
User-agent: wxyz
Disallow:   /

User-agent: IBM-WebCrawler
Disallow:   /htmldocs/temp
```

Глава 6. Введение в рабочие потоки

Рабочий поток EIP используется для управления ходом и производительностью работы в организации. Пользователям, работающим с результатами объединенного поиска, часто надо решать, какие действия им нужно выполнить. Рабочий поток EIP можно использовать, чтобы заранее определить, как пользователи могут выполнять работу.

Рабочие потоки можно автоматизировать, настроив профили и правила, управляющие взаимодействием компонентов рабочего потока. Кроме того, можно выбирать уровень ограничений системы, управляя доступом и правами пользователей при помощи наборов привилегий и списков управления доступом.

Что такое рабочий поток

Большинство деловых операций можно охарактеризовать как набор взаимосвязанных процессов. Работа передается от одного сотрудника к другому и из одного отдела в другой. Некоторые простые процессы требуют всего несколько шагов, для более сложных процессов требуется много сотрудников из разных отделов.

Рабочий поток позволяет перемещать работу по процессу и принимать решения относительно работы на протяжении всего этого процессе. В нашем примере компания XYZ Insurance получает по почте большое количество страховых исков. В процессе проверки оценщики страховых исков должны собрать документы, например, фотографии, оценки ущерба и отчеты экспертов. Каждый день в течение нескольких часов сотрудники открывают, сортируют, подшивают и отслеживают информацию, а также собирают документы, относящиеся к делу, для окончательного рассмотрения.

По мере получения и проверки эта информация перемещается от одного сотрудника к другому. Когда информация по иску собрана, она может обрабатываться сотрудниками сразу в нескольких отделах.

Как использовать рабочий поток

Большинство предприятий, обрабатывающих документы (как в приведенном примере XYZ Insurance), выполняют некоторые (или все) из следующих задач:

- Подшивают документы для последующих требований.
- Собирают документы, бланки, отчеты и информацию из разных источников, после чего направляют эти документы на обработку.
- Сравнивают входящую почту с обрабатываемыми документами.

Рабочий поток отражает этот поток работы. Он описывает действия, которые можно выполнить над группой из одного или нескольких документов или содержимое и путь, который эта группа документов проходит по рабочему потоку. Рабочий поток отражает путь выполнения работы с ясно определенной областью действия и границами. Он определяет последовательность рабочих этапов и заданий, а также связи и отношения между рабочими этапами и заданиями. Рабочий поток определяет критерии, используемые для принятия решений о потоке работы. Информацию о процессе создания рабочего потока смотрите в книге *Workstation Application Programming Guide*. Информацию об использовании клиента с рабочим потоком смотрите в руководстве *eClient. Установка, конфигурирование и управление*.

Синхронизация ID пользователей и групп

В этом разделе объясняется, как синхронизировать ID пользователей и группы между EIP, Content Manager и MQ Series Workflow.

Если вы управляете пользователями в Content Manager или Enterprise Information Portal в системе, содержащей MQSeries Workflow, вы также управляете пользователями в MQSeries Workflow. Поэтому создавая, изменяя или удаляя ID пользователя или группу в Content Manager или Enterprise Information Portal, нужно затем сделать то же самое и для сервера MQSeries Workflow.

Поскольку Content Manager и Enterprise Information Portal совместно используют одни и те же ID пользователей и группы, если создать ID пользователя или группу в клиенте администратора системы при включенной службе рабочего потока, но не запущенном сервере MQSeries Workflow, будет выдано сообщение об ошибке, указывающее, что этот пользователь или группа не были созданы на сервере MQSeries Workflow. Этот ID пользователя будет существовать как пользователь в Content Manager или Enterprise Information Portal, но не будет существовать на сервере MQSeries Workflow.

Для синхронизации ID пользователей и групп для управления содержимым с ID пользователей и групп для MQSeries Workflow, нужно запустить утилиту синхронизации пользователей Workflow. Если сервер MQSeries Workflow установлен вручную, убедитесь, что он запущен. Откройте Службы и проверьте состояние MQSeries Workflow *X.X* - FMC (где *X.X* - версия установленного Workflow). Если MQSeries Workflow не запущен, запустите эту службу, или, если установка выполнялась в автоматическом режиме, перейдите в каталог WFInstall и запустите пакетный файл `CMBWFStart.bat`. Для запуска утилиты синхронизации:

1. Перейдите в каталог, в котором установлен Enterprise Information Portal. Каталог по умолчанию - `C:\CMBROOT`.
2. Введите `EIPUser2WF.bat`.

   Внимание: Если у вас удаленный сервер рабочего потока, перед вводом команды `EIPUser2WF.bat` нужно запустить пакетные файлы или файлы оболочки сервера RMI (`cmbsvregist81.bat` или `cmbsvregist81.sh`). Кроме того, чтобы найти локальный или удаленный сервер, `EIPUser2WF.bat` использует информацию из двух INI-файлов: `cmbsvcs.ini` и `cmbsvclient.ini`. Для локального сервера в `cmbsvcs.ini` должно быть указано LOCAL, а для удаленного - REMOTE. Файл `cmbsvclient.ini` содержит информацию о положении сервера.
3. Введите требуемую информацию для имени базы данных EIP, ID пользователя, пароля и схемы.

Когда вы введете необходимую информацию, утилита синхронизации скопирует всех пользователей с сервера Content Manager или Enterprise Information Portal на сервер MQSeries Workflow. Когда утилита завершит работу, вы не будете получать сообщений о том, что пользователь или группа неизвестны на сервере MQSeries Workflow.

Удаляя ID пользователя или группу из Content Manager или Enterprise Information Portal, нужно также удалить этот ID пользователя или группу с сервера MQSeries Workflow. Если ID пользователя или группа не существуют на сервере MQSeries Workflow, вы не сможете удалить их из Content Manager или Enterprise Information Portal. Например, вы создаете ID пользователя или группу в Content Manager или Enterprise Information Portal при выключенной опции сервера рабочего потока EIP. Затем вы включаете эту опцию сервера рабочего потока и пытаетесь удалить этот ID пользователя или группу. Поскольку этот ID пользователя или группа не существует

в MQSeries Workflow, вы получите сообщение об ошибке, указывающее, что этот ID пользователя не существует на сервере MQSeries Workflow. Чтобы исправить эту ошибку, нужно запустить утилиту EIPUser2WF для синхронизации ID пользователей и групп и затем удалить этот ID пользователя или группу.

Переустановка сервера EIP с включенным рабочим потоком

Если на сервере EIP разрешен рабочий поток EIP и вы хотите установить новый сервер EIP, необходимо удалить все данные рабочего потока EIP с сервера MQSeries Workflow.

Перед удалением сервера EIP выполните следующие действия для удаления всех данных рабочего потока EIP с сервера MQSeries Workflow. **Внимание:** Выполняйте эти действия в указанном ниже порядке.

1. Завершите все экземпляры рабочего потока, используя eClient или API рабочего потока. При завершении экземпляров рабочего потока удаляются все рабочие элементы.
2. Удалите все рабочие списки и шаблоны рабочих потоков, используя клиент администратора системы EIP или API управления системой.
3. Удалите всех пользователей и все группы из EIP при включенном рабочем потоке, чтобы все эти ID пользователей и группы пользователей были также удалены и с сервера MQSeries Workflow.
4. Переконфигурируйте сервер MQSeries Workflow для EIP:
   a. Вставьте установочный компакт-диск EIP. В командной строке перейдите в каталог `WfInstall`.
   b. В командной строке введите:

      ```
      fmcibie -iCMBWFAdmin.fdl  -uadmin -ppassword -o  -f
      ```

Внимание: Если удалить базу данных EIP перед удалением ID пользователей и групп из EIP и MQSeries Workflow и затем попытаться создать те же ID пользователей и группы, что были в удаленной базе данных EIP, вы получите сообщение об ошибке, указывающее, что эти пользователи и группы не могут быть добавлены в EIP. Это сообщение об ошибке: DGL2616A: Не удалось добавить пользователя: XXX -DGL2485A: Этот пользователь рабочего потока уже существует. Чтобы исправить эту ошибку:

1. Выключите службу рабочего потока EIP в клиенте администратора системы EIP.
2. Выйдите из клиента администратора системы EIP и заново зарегистрируйтесь в нем. Зарегистрировавшись после выключения службы рабочего потока, вы сможете создать те же ID пользователей и группы, которые существовали в EIP и все еще существуют на сервере MQSeries Workflow.
3. Создав ID пользователей и группы, уже существующие на сервере MQSeries Workflow, включите службу рабочего потока EIP.

Синхронизация ID пользователей и групп между MQSeries Workflow и базой данных EIP

Для синхронизации ID пользователей и групп, существующих на сервере MQSeries Workflow, но не существующих в базе данных EIP, можно сделать следующее:

1. Создайте файл и введите в него такой текст:
   - CODEPAGE 1252
   - FM_RELEASE V3R3 2
   - DELETE PERSON ’User1’

- DELETE PERSON 'User2'
- DELETE PERSON 'User3'

где User1, User2 и User3 - пользователи, которых нужно удалить. Можно указать любое нужное число пользователей.

2. Сохраните этот файл и выполните из командной строки такую команду:

```
fmcibie -u admin -ppassword -i DeletePersons.fdl  -f -o
```

где `DeletePersons.fdl` - имя файла, созданного на предыдущем шаге.

Для групп, существующих на сервере MQSeries Workflow, но более не существующих в базе данных EIP, можно сделать следующее:

1. Создайте текстовый файл и введите в него такую информацию:
   - CODEPAGE 1252
   - FM_RELEASE V3R3 2
   - DELETE ROLE 'Group1'
   - DELETE ROLE 'Group2'
   - DELETE ROLE 'Group3'

   где Group1, Group2 и Group3 - группы, которые нужно удалить. Можно указать любое нужное число групп.
2. Сохраните этот файл и введите в командной строке такую команду:

```
fmcibie -u admin -p password -i DeleteGroups.fdl  -f -o
```

где `DeleteGroups.fdl` - имя файла, созданного на предыдущем шаге.

При удалении ID пользователя из EIP при включенной поддержке рабочего потока вы можете получить такое сообщение об ошибке:

```
Не удалось удалить пользователя [RC=12]
```

При ошибке удаления пользователя система создает файл `temp.log`, содержащий подробную информацию об ошибке. Файл `temp.log` записывается в каталог *x:*`\CMBROOT`. **Подсказка:** При попытке удалить ID пользователя, под которым вы зарегистрированы в клиенте MQSeries Workflow, возникнет ошибка операции удаления.

Планирование рабочего потока

Перед тем, как начать определять рабочий поток, надо проанализировать выполняемую работу, где, как и кем она выполняется. Шаг планирования выполняет администратор или бизнес-аналитик.

Что представляет собой конечный продукт? Конечным продуктом может быть результат всей работы, выполненной предприятием, одним из отделов предприятия или отдельными сотрудниками из разных отделов. Например, конечный продукт процесса обработки исков в страховой компании XYZ Insurance - письмо владельцу полиса, принимающее или отклоняющее иск.

Проанализируйте информацию, которую нужно обработать для создания конечного продукта, определите, какие действия нужно выполнить и когда их выполнять, и решите, как информация должна проходить по рабочему потоку.

Информация для обработки

Рассмотрите информацию, которую должны обрабатывать пользователи в вашей организации. Какие типы входных данных необходимы для конечного продукта? Какие конкретно документы нужно обработать?

Рабочий элемент - это не сам документ. *Рабочий элемент* содержит ссылку на документ и другую информацию о документе (такую, как состояние документа, дата создания и т.п.). *Рабочий элемент* может быть любыми данными (документами или объектами) с контент-сервера. Например, компания XYZ Insurance сначала получает бланки исков, а потом - сопроводительные документы, например, фотографии, оценки ущерба и отчеты экспертов.

Как обрабатывается информация

Кто может лучше всех справиться с каждым шагом процесса? Например, помощник администратора может проверить, заполнен ли бланк иска, после чего положить его в папку до тех пор, пока от владельца полиса не придет определенный документ. По получении документа оценщик исков отвечает за сличение документа с бланком иска и за принятие этого документа в работу.

Можно считать *рабочий список* создаваемой вами очередью работ для одного или нескольких сотрудников. Бланки исков можно сгруппировать в *рабочий список*, доступный некоторому числу оценщиков исков. Рабочий список - это фильтрованное представление рабочих элементов. Сотрудники видят только те элементы рабочего списка, которые им разрешено видеть.

Рабочие списки можно определять так, чтобы управлять всеми деталями обработки исков, например, сбором фотографий, оценок ущерба и отчетов. Кроме того, рабочий список может включать в себя работу из разных рабочих потоков. Например, рабочий список для одного из оценщиков исков может содержать оценки ущерба для одного иска, фотографии для другого иска и отчет эксперта по третьему иску. Для каждого элемента рабочего списка оценщик может выполнять разные действия. Он может изучить оценку ущерба и принять первый иск. Прежде чем взять в работу по второму иску фотографии, он может решить дождаться получения дополнительной информации. Для третьего иска он может переслать отчет эксперта другому сотруднику, чтобы он предпринял необходимые действия.

Выполняемые действия

Решите, какие действия нужно выполнить с содержимым рабочего элемента в рабочем потоке. Например, оценщик исков может принять страховой иск или отклонить его, если он недостаточно полон. *Список действий* определяет, какие действия разрешено выполнять пользователю над работой.

Например, в зависимости от того, отвечает ли иск начальным требованиям, оценщик может выбрать продолжение работы с иском в рабочем потоке или отклонение этого иска.

Как информация перемещается по процессу

Решите, как должна перемещаться информация и какие рабочие этапы она должна проходить. Например, когда нужно пересматривать страховой иск? Какие сопроводительные документы нужны для перехода к следующему шагу процесса? Какими критериями определяется, принять или отклонить иск? Этот поток информации является основой рабочего потока.

Рабочий поток состоит из путей, по которым движется работа во время выполнения. Откуда берутся входные данные? У рабочего потока должна быть некоторая начальная точка. Для компании XYZ Insurance документом, с которого начинается рабочий поток, является исковое заявление, полученное от владельца полиса.

Когда все документы получены, рабочий элемент может продолжать свой путь к конечному действию – например, к принятию иска.

Как связать все это вместе

Если вы проанализировали информацию, которую нужно обработать, решили, какие действия нужно выполнить, и выбрали путь прохождения информации, вы готовы к составлению диаграммы рабочего потока, то есть графического представления вашего рабочего потока. Для создания диаграммы используется построитель рабочего потока EIP.

На диаграмме рабочего потока показано, как работа перемещается по рабочим этапам процесса, и отмечены задачи, входящие в каждый рабочий этап. Она описывает поток, основные элементы и ключевые точки рабочего потока.

Каждый символ на диаграмме рабочего потока обозначает точку, в которой выполняется работа. Страховой иск нужно изучить, сопроводительную документацию нужно собрать, а иск нужно в зависимости от определенных критериев принять или отклонить. Дополнительную информацию об используемых в построителе рабочего потока символах процесса смотрите в разделе “Создание рабочего потока” на стр. 102.

Использование компонентов рабочего потока Enterprise Information Portal

В этом разделе описываются компоненты рабочего потока. Доступ ко всем компонентам осуществляется через клиент администратора. **Совет:** в рабочий поток EIP Версии 8 внесены некоторые изменениями, включая изменения, внесенные в контейнер Версии 7.1, для соответствия новой архитектуре Content Manager Версии 8.

Использование построителя рабочего потока

Построитель рабочего потока используется для графического определения и построения рабочего потока для рабочей группы, отдела или целой организации. **Ограничение:** процесс перенастройки EIP перенастраивает пользователей из баз данных Версии 7.1. В EIP Версии 8.2 не предусмотрена автоматическая перенастройка данных рабочего потока. Ваши диаграммы рабочего потока из Версии 7.1 необходимо перерисовать с помощью построителя рабочего потока EIP Версии 8.2 и перезапустить процессы рабочего потока из EIP Версии 7.1.

Перед тем, как при помощи построителя рабочего потока создавать рабочие потоки, надо определить наборы привилегий, списки управления доступом, пользователей, группы пользователей, действия, списки действий и рабочие списки. При определении рабочего потока в клиенте администратора можно задать список действий по умолчанию для всего рабочего потока. Можно также назначить разные списки действий разным узлам рабочего потока. Дополнительную информацию об этих задачах смотрите в разделах “Определение списков действий” на стр. 102, “Определение рабочих списков” на стр. 101 и в электронной справке.

Хотя построитель рабочего потока используется для построения рабочих потоков, при его помощи нельзя запустить рабочий поток. При помощи клиента пользователи

смогут просматривать рабочие списки и рабочие элементы и выполнять над ними действия. Дополнительную информацию о том, как создать собственный клиент для работы с рабочим потоком Enterprise Information Portal, смотрите в книге *Workstation Application Programming Guide* и в электронном справочнике по API. В Табл. 7 показаны три значка, общие для всех рабочих потоков. Существуют и другие значки рабочих потоков, такие как подпоток, событие, узел сбора и обработчик пользователя. В электронной справке по EIP дается подробное описание значков панели инструментов.

Таблица 7. Основные значки рабочих потоков

Значок	Описание
	С начального узла начинается процесс рабочего потока. В диаграмме процесса рабочего потока должен быть один и только один начальный узел.
	Конечный узел заканчивает процесс рабочего потока. Каждая новая диаграмма процесса рабочего потока содержит конечный узел. При создании процесса конечный узел генерируется автоматически. Его можно переместить в любую точку холста. В диаграмме процесса рабочего потока должен быть один и только один конечный узел.
	Рабочий узел связывает рабочий список и список действий для конкретной точки процесса рабочего потока. Рабочий узел представляет точку процесса рабочего потока, в которой выполняется работа. Для каждого узла, включая начальный узел и конечный узел, должен быть задан список действий и должно быть указано, кто является исполнителем этого набора действий.

Использование служб рабочих потоков

Enterprise Information Portal предоставляет службы рабочих потоков, управляющие информацией рабочих потоков. Рабочие потоки и списки действий, созданные при помощи построителя рабочего потока, хранятся в базе данных управления Enterprise Information Portal и базе данных рабочих потоков IBM MQSeries.

После создания администратором системы рабочего списка информация, связанная с этим рабочим списком, постоянно хранится в управляющей базе данных. Администратор системы может изменять, удалять и добавлять рабочие списки при помощи EIP. Когда администратор системы резервирует рабочий поток, этот рабочий поток блокируется в базе данных Enterprise Information Portal и помечается в ней как зарезервированный пользователем, что предотвращает изменение рабочего потока другими пользователями, пока этот пользователь не закончит работу с ним.

Определение рабочих списков

Рабочий список можно представить как выборку из всей имеющейся работы. Рабочий список - это фильтрованный список элементов, назначенный отдельным пользователям или группам пользователей. Когда пользователи регистрируются в Enterprise Information Portal, они видят фильтрованные списки назначенных им рабочих элементов. Для определения рабочих списков используется клиент администратора Enterprise Information Portal.

В определение рабочего списка входят правила, управляющие выводом, состоянием и защитой его рабочих элементов. Эти правила задаются для каждого рабочего списка во время его создания. Для управления доступом к рабочему списку нужно создать

для этого рабочего списка список управления доступом. Полное описание того, как определяются рабочие списки, смотрите в электронной справке. Определение рабочего списка включает в себя:

Список управления доступом
Список управления доступом содержит один или несколько ID отдельных пользователей или групп пользователей и набор привилегий, связанный с каждым из них. Этот набор привилегий используется для определения авторизации пользователя, необходимой для доступа к определенным рабочим заданиям или для их выполнения. Списки управления доступом применяются для ограничения доступа пользователей к элементам рабочего списка.

Фильтрование и сортировка рабочих списков
Критерии, по которым пользователю доступен для просмотра фильтрованный и отсортированный рабочий список.

Максимальное число элементов в рабочем списке
Максимальное число элементов, заданное вами для рабочего списка.

Определение списков действий

Список действий - это исчерпывающий список действий, которые разрешено выполнять пользователю над работой в рабочем потоке.

Пошаговые инструкции по определению действий и списков действий смотрите в электронной справке клиента администратора.

Создание рабочего потока

После определения действий, списков действий и рабочих списков нужно при помощи построителя рабочего потока создать модель рабочего потока. Пошаговые инструкции по определению действий и списков действий смотрите в электронной справке клиента администратора. В построителе рабочего потока предусмотрены визуальные подсказки для создания рабочего потока.

Включение построителя рабочего потока

На этом шаге вы запускаете рабочий поток в управляющей базе данных.
Ограничение: база данных, выбранная для рабочего потока, должна находиться на том же сервере, на котором установлен MQ Series, а службы MQSeries должны быть запущены.

Чтобы разрешить рабочий поток EIP и создать определение рабочего потока:

1. Зарегистрируйтесь на клиенте администратора.
2. Если управляющих баз данных несколько, щелкните по значку базы данных, где вы хотите разрешить рабочий поток.
3. Выберите Инструменты -> Службы. Нажмите кнопку Разрешить рабочий поток.
4. Выйдите из клиента и зарегистрируйтесь на нем снова. Если баз данных несколько, выберите значок базы данных, где запущен рабочий поток. Появится значок папки **Рабочие потоки**.
5. На левой панели главного окна управления Enterprise Information Portal дважды щелкните по папке **Рабочие потоки**.
6. Щелкните правой кнопкой мыши по значку Определения рабочих потоков и выберите **Новое**, чтобы создать определение рабочего потока.

Обязательно: Перед определением рабочего потока надо создать как минимум один список управления доступом, одно действие и один список действий.

Запуск сервера MQSeries Workflow

Чтобы запустить сервер MQSeries Workflow, введите в командной строке команду `cmbwfstart`. Откроются два окна сервера MQSeries Workflow. Оставьте эти командные окна открытыми, чтобы сервер продолжал работать.

Если рабочий поток установлен после начальной установки Enterprise Information Portal, надо сконфигурировать рабочие потоки в системе Enterprise Information Portal. Кроме того, конфигурацию надо изменить, если рабочие потоки установлены не на ту рабочую станцию, на которой установлен клиент администратора.

1. В окне управления щелкните по файловому элементу **Инструменты**.
2. В меню выберите **Службы**.
3. Включите переключатель **Рабочий поток**.
4. После завершения конфигурации для инициализации рабочих потоков выйдите из клиента администратора Enterprise Information Portal и зарегистрируйтесь снова. После регистрации в клиенте администратора Enterprise Information Portal на левой панели появится значок **Определения рабочего потока**.

Совет: Администраторы не увидят значок **Определения рабочих потоков**, пока не получат право управления возможностью рабочего потока. Информацию об ограничении доступа к рабочим потокам смотрите в руководствах администраторов систем для соответствующих контент-серверов. Дополнительную информацию о предоставлении администраторам прав для управления рабочими потоками смотрите в электронной справке.

Клиент можно создать из пользовательской прикладной программы при помощи набора инструментов и примеров соединителя EIP; можно также воспользоваться примером клиента EIP.

Глава 7. Файлы примеров IBM Web Crawler

В этом разделе приводится два примера кода. Файл примера config-sample2.xml содержит примеры параметров конфигурации <gcs-config>. Пример анализа журнала содержит отчета с информацией о выполненном процессе искателя.

Пример config-sample2.xml

В этом разделе приводится файл примера gcs-config.

```
<!DOCTYPE gcs-config SYSTEM "config.dtd">
<gcs-config>
   <!-- Глобальные параметры: -->
   <globals max-urls="1000000"
            num-crawlers="30"
            num-summarizers="8"
            summaries-dir="summaries"
            log-file="log/LOG.txt"
            temp-dir="temp"
            log-priority="warn"
            text-monitor="60"
            graph-monitor="2"
            connect-timeout="120"
            read-timeout="100">

      <!-- параметры записи в журнал -->
      <logger-config category="gcs.summaries.list.resource"
priority="info" log-file="log/resources.txt"/>
      <logger-config category="gcs.summaries.list.host"     priority="info"
log-file="log/hosts.txt"/>
      <logger-config category="gcs.url.skipped"             priority="info"
log-file="log/skipped_urls.txt"/>
      <logger-config category="gcs.url.fromto"              priority="info"
log-layout="%m\n" log-file="log/fromto.txt"/>
      <logger-config category="gcs.http"                    priority="info"
log-file="log/http.txt"/>
      <logger-config category="gcs.http.connect"            priority="info"
log-file="log/connecterrs.txt"/>

      <!-- задание базы данных
      <urlpool-config urlcontainer-class="DB2URLContainer"
urlcollection- class="DB2URLCollection">
         <urlpool-param name="dbname"    value="gcs"/>
         <urlpool-param name="user"      value="xxxxxx"/>
         <urlpool-param name="password"  value="xxxxxx"/>
         <urlpool-param name="cachesize" value="1000"/>
         <urlpool-param name="driver"
value="COM.ibm.db2.jdbc.app.DB2Driver"/>
      </urlpool-config> -->

      <!-- задание прокси SOCKS
      <system-properties>
         <property name="socksProxySet" value="true"/>
         <property name="socksProxyHost" value="socks2.server.ibm.com"/>
         <property name="socksProxyPort" value="1080"/>
      </system-properties> -->

   </globals>

   <group-list>
      <group name="ibm">
```

```
          <crawler-config recursion-depth="-1">
             <seed-list>
             <!-- Начальные URL процесса искателя: -->
                <seed url="http://gcs.stl.ibm.com/gcs/testurl.html"/>
                <seed url="http://gcs.stl.ibm.com/gcs/stl.html"/>
                <seed url="http://gcs.stl.ibm.com/gcs/ibm.html"/>
             </seed-list>

             <content-type-pattern-list>
             <!-- Файлы с расширениями, не перечисленными здесь,
                  рассматриваться не будут: -->
                <url-name-pattern name="htm*"/>
                <url-name-pattern name="pdf"/>
                <url-name-pattern name="gif"/>
                <url-name-pattern name="zip"/>
                <url-name-pattern name="txt"/>
             </content-type-pattern-list>

             <include-pattern-list>
             <!-- URL, не соответствующие перечисленным шаблонам,
                     рассматриваться не будут: -->
                <url-obj-pattern host="*.ibm.com"/>
       <!-- sbo - url-obj-pattern   query="*OpenDocument*"/ -->
       <!-- sbo - url-obj-pattern   query="*OpenView*"/ -->
                <!-- url-obj-pattern   query="*OpenDocument =>
OpenDocument&amp;ExpandAll*"/ -->
       <!-- url-obj-pattern   query="*OpenView =>
OpenView&amp;ExpandAll&amp;Count=999999*"/ -->
             </include-pattern-list>

             <exclude-pattern-list>
             <!-- URL, соответствующие перечисленным шаблонам,
                     рассматриваться не будут: -->
                <!-- пропускаем следующие URL в вашей интрасети -->
                <url-obj-pattern file="*news*"/>
                <url-obj-pattern file="*search*"/>
                <url-obj-pattern file="*/afs/*"/>
                <url-obj-pattern file="*/.../*"/>
                <url-obj-pattern file="*bluepages*"/>
                <!-- пропускаем личные домашние страницы -->
                <url-obj-pattern file="*/~*"/>
                <!-- пропускаем SOCKS: в URL не должно быть указано прямо -->
                <url-regex-pattern match=".*:1080/.*"/>
                <!-- пропускаем шлюзы: рекомендуется для большинства -->
                <url-regex-pattern match=".*[\?\=\+\;\%&quot;&amp;].*"/>
                <!-- иначе просматриваем шлюзы, как сконфигурировано... -->
                  <!-- пропускаем Domino -->
                  <url-obj-pattern   file="*.nsf*"/>
                  <!-- иначе просматриваем Domino:
                          допускается только OpenDocument -->
                  <url-obj-pattern   query="*OpenServer*"/>
                  <url-obj-pattern   query="*OpenDatabase*"/>
                  <url-obj-pattern   query="*OpenElement*"/>
                  <url-obj-pattern   query="*OpenView*"/>
                  <url-obj-pattern   query="*OpenAbout*"/>
                  <url-obj-pattern   query="*OpenHelp*"/>
                  <url-obj-pattern   query="*OpenIcon*"/>
                  <url-obj-pattern   query="*OpenForm*"/>
                  <url-obj-pattern   query="*OpenNavigator*"/>
                  <url-obj-pattern   query="*OpenAgent*"/>
                  <url-obj-pattern   query="*CreateDocument*"/>
                  <url-obj-pattern   query="*DeleteDocument*"/>
                  <url-obj-pattern   query="*EditDocument*"/>
                  <url-obj-pattern   query="*SaveDocument*"/>
                  <url-obj-pattern   query="*SearchSite*"/>
                  <url-obj-pattern   query="*SearchView*"/>
                  <url-obj-pattern   query="*&amp;login*"/>
```

```
                <url-obj-pattern   query="*Command*"/>
                <!-- просматриваем Domino: избегаем изменений OpenDocument -->
                <url-obj-pattern   query="*ExpandSection*"/>
                <url-obj-pattern   query="*Navigate*"/>
                <url-obj-pattern   query="*Start*"/>
             <!-- -->

          </exclude-pattern-list>
        </crawler-config>

        <summarizer-config>
        <!-- Составитель сводок по умолчанию - Copy2Rdf.
                Для следующих типов используется: -->

          <resource-handler content-type="*htm*"
                            summarizable="EipHtmlSummarizable"
                            summary-maker="EipHtmlRawSummaryMaker" />
          <resource-handler content-type="*pdf"
                            summarizable="InsoSummarizable"
                            summary-maker="InsoSummaryMaker" />
        </summarizer-config>
v     </group>
   </group-list>
</gcs-config>
```

Пример анализа файла журнала IBM Web Crawler

```
D:\gcs\run\log>perl loganalysis.pl log.txt

Затраченное время файла Log.txt на 7710 строк - 1,84 минуты.

GCS was configured for 20 crawlers
        999  total crawls attempted
        137  - total crawl failures:
                  21    GCSHttpConnection.ABANDONING
                  12    GCSHttpConnection.CONNECT_ERROR
                  16    GCSHttpConnection.UNKNOWN_HOST
                   4    HTTP 403
                  29    HTTP 404
                   2    HTTP 500
                   8    HTTP 599
                   1    Read timed out
                  39    Robots not allowed
                   4    over max redirects
                   1    unknown protocol
   --------
        862  = successfully crawled
          0  - unchanged since earlier crawl
   --------
        862  = new or changed
        468  crawled per minute


GCS was configured for 5 summarizers
        855  total summaries attempted
          0  - total summary failures:
   --------
        855  = successfully summarized
        144  gcs.summaries.list.host
        855  gcs.summaries.list.resource
        465  summarized per minute


GCS successfully crawled 134 servers to obtain 862 URL:
  afqa0854.mop.ibm.com: 15
  als1f1.yamato.ibm.com: 1
```

```
apache.btv.ibm.com: 1
apc.endicott.ibm.com: 2
as400service.ibm.com: 1
atlas.bocaraton.ibm.com: 1
autoproxy.ibm.com: 1
cer.si.ibm.com: 1
commerce.www.ibm.com: 1
crmweb.boulder.ibm.com: 3
d02ntcl01.ibm.com: 1
dacs.endicott.ibm.com: 1
duke.toraix.can.ibm.com: 1
ebcweb.austin.ibm.com: 1
ecspubs.ibmus2.ibm.com: 5
edaw3.fishkill.ibm.com: 1
endwww.endicott.ibm.com: 1
gcs.stl.ibm.com: 1
gustwick.austin.ibm.com: 1
ibmfnsys.somers.hqregion.ibm.com: 1
ibmpnyi1.somers.hqregion.ibm.com: 2
ifw-www.mul.ie.ibm.com: 1
iplswww.nas.ibm.com: 2
itirc.ibm.com: 1
logosite.services.ibm.com: 1
lt.lahulpe.ibm.com: 17
messaging.ibm.com: 1
mrsmrn04.leeds.uk.ibm.com: 1
online.lahulpe.ibm.com: 1
page.sg.ibm.com: 1
procure.sby1.ibm.com: 1
reso.somers.hqregion.ibm.com: 1
riscta1.leipzig.de.ibm.com: 1
rrhhar.argentina.ibm.com: 1
seashore.stl.ibm.com: 1
secureway.raleigh.ibm.com: 15
service.software.ibm.com: 1
software.ibmus2.ibm.com: 1
techcenter.austin.ibm.com: 1
tr2.fishkill.ibm.com:8080: 1
ucd.torolab.ibm.com: 1
usmweb.boulder.ibm.com: 1
w3-1.ibm.com: 32
w3-2.ibm.com: 3
w3-3.ibm.com: 108
w3-5.ibm.com: 4
w3.a-nz.au.ibm.com: 1
w3.academy.ibm.com: 1
w3.almaden.ibm.com: 2
w3.alphaworks.ibm.com: 1
w3.ap.ibm.com: 1
w3.asca.ibm.com: 7
w3.austin.ibm.com: 3
w3.boulder.ibm.com: 1
w3.br.ibm.com: 1
w3.btv.ibm.com: 1
w3.can.ibm.com: 40
w3.chq.ibm.com: 4
w3.coc.ibm.com: 1
w3.corporatetechnology.ibm.com: 1
w3.cupertino.ibm.com: 1
w3.dds.dfw.ibm.com: 17
w3.demopkg.ibm.com: 4
w3.design.ibm.com: 1
w3.developer.ibm.com: 3
w3.education.ibm.com: 1
w3.emea.ibm.com: 14
w3.enterlib.ibm.com: 7
w3.finsys.ibm.com: 1
```

```
w3.gcg.ibm.com: 1
w3.globalfinancing.de.ibm.com: 1
w3.hakozaki.ibm.com: 1
w3.houston.ibm.com: 1
w3.hursley.ibm.com: 5
w3.iabc.ibm.com: 1
w3.ibm.com: 180
w3.ibmfax.ibm.com: 1
w3.ibmla.ibm.com: 14
w3.isicc.de.ibm.com: 1
w3.itso.ibm.com: 1
w3.japan.ibm.com: 1
w3.knowledge.raleigh.ibm.com: 1
w3.linux.ibm.com: 1
w3.marketiq.ibm.com: 1
w3.micro.ibm.com: 2
w3.mtlisc.can.ibm.com: 1
w3.munich.ibm.com: 1
w3.ode.raleigh.ibm.com: 1
w3.paylink.au.ibm.com: 1
w3.pisc.uk.ibm.com: 1
w3.pl.ibm.com: 1
w3.printers.ibm.com: 1
w3.pssc.mop.ibm.com: 1
w3.pssed.au.ibm.com: 1
w3.raleigh.ibm.com: 3
w3.rchland.ibm.com: 1
w3.research.ibm.com: 3
w3.reserve.ibm.com: 1
w3.rs6000.ibm.com: 1
w3.security.ibm.com: 1
w3.software.ibm.com: 6
w3.ssd.ibm.com: 1
w3.stl.ibm.com: 1
w3.techline.ibm.com: 1
w3.techsupp.yamato.ibm.com: 1
w3.torolab.ibm.com: 2
w3.usergroup.ibm.com: 1
w3.vendor.pok.ibm.com: 1
w3.viewblue.ibm.com: 1
w3.watson.ibm.com: 2
w3.wdg.uk.ibm.com: 1
w3.ytal.yasu.ibm.com: 1
w3.zurich.ibm.com: 1
w3chq.disbursements.ibm.com: 1
w3is.lagaude.ibm.com: 1
w3md.btv.ibm.com: 1
w3ssd.mainz.de.ibm.com: 1
w3vm.demopkg.ibm.com: 1
widweb.raleigh.ibm.com: 1
wtscpok.itso.ibm.com: 1
wwas.raleigh.ibm.com: 1
www-1.ibm.com: 63
www-3.ibm.com: 4
www-4.ibm.com: 86
www.almaden.ibm.com: 1
www.as400.ibm.com: 1
www.chips.ibm.com: 1
www.ibm.com: 52
www.ieg.ibm.com: 1
www.patents.ibm.com: 1
www.pc.ibm.com: 2
www.rs6000.ibm.com: 23
www.software.ibm.com: 9
www.storage.ibm.com: 1
www.watson.ibm.com: 1
```

```
GCS timed out 1 times:
    w3-3.ibm.com: 1

GCS ignored 42 URL prohibited by robots.txt:
  reso.somers.hqregion.ibm.com: 1
  w3.education.ibm.com: 1
  w3.rchland.ibm.com: 34
  w3.zurich.ibm.com: 1
  www.ibm.com: 5

GCS skipped 3846 URL (requires gcs.url logging)
  59 specified an unsupported protocol:
    protocol not supported  gopher: 6
    protocol not supported  mailto: 53
  1206 had content-types (lower or UPPER case, > 10) that were not included
    .2: 12
    .faq: 13
    .1: 14
    .asp: 16
    .cgi: 21
    .shtml: 90
    .pl: 92
    .gif: 157
    .nsf: 160
    .jpg: 214
    .css: 240
  516 URL were on servers and/or paths that were not included
  2065 were excluded for these reasons:
    URL longer than 254: 1
    excluded by rule 1: 1210
    excluded by rule 2: 854
```

Глава 8. Использование текстового поиска и QBIC

В первом разделе этого приложения описывается конфигурирование и использование текстового поиска и запроса по содержимому изображения (QBIC), двух возможностей, которые доступны, если во время установки EIP выбран соединитель Content Magnager Версии 7.1. Во втором разделе этого приложения содержится информация о загрузке примеров текстов и изображений, используемых с примерами прикладных программ.

Поиск документов при помощи механизма текстового поиска

Текстовый поиск можно включить в сервер Content Manager Версии 7.1, в результате вы сможете автоматически индексировать, искать и получать документы, хранимые в Content Manager. Пользователи могут искать документы по словам или словосочетаниям. Сервер текстового поиска поддерживает как однобайтные, так и двухбайтные наборы символов и работает под AIX и Windows.

В текстовый поиск включена поддержка структурированных документов XML, HTML и документов в формате ASCII с тегами, что дает возможность искать термины внутри заданных разделов документов. Можно искать данные во вложенных разделах. Можно выполнять поиск по полному контексту XML, например, искать ″IBM″ во всех заголовках или же в заголовке определенного раздела. Если задать путь к DTD (document tag definition - определение тегов документов), текстовый поиск сможет использовать соответствующее DTD динамически для каждого документа, если ссылка на DTD хранится как метаданные для этого документа.

Информацию о планировании и установке системы EIP с поддержкой текстового поиска смотрите в книге *Планирование и установка Enterprise Information Portal*.

Включение сервера текстового поиска

Чтобы использовать сервер текстового поиска, перед запуском клиента администратора IBM Content Manager for Multiplatforms надо разрешить управление этим сервером. Чтобы разрешить управление:

1. Запустите библиотечный сервер IBM Content Manager for Multiplatforms.

   Дайте библиотечному серверу выполнить построение индексных классов.
2. Запустите сервер текстового поиска на том компьютере, где он установлен, введя команду:

   ```
   imlss -start экземпляр
   ```

   где *экземпляр* - имя экземпляра сервера текстового поиска, выбранное при установке или заданное при помощи командной утилиты *imlcfgsv*.

Поиск изображений при помощи запроса по содержимому изображения (QBIC)

В этом разделе вы ознакомитесь с запросом по содержимому изображения (QBIC), его конфигурированием и использованием. Возможность QBIC доступна, только если вы установили соединитель Content Manager Версии 7.1. QBIC совместим с операционными системами Windows и AIX.

Введение в поиск изображений

Сервер поиска изображений, используя разработанную IBM технологию QBIC (запрос по содержимому изображения), позволяет искать объекты по их видимым свойствам, например, по цвету и текстуре. Сервер поиска изображений анализирует изображения и сохраняет информацию о них в базе данных. После этого пользователи могут строить запросы, где используются видимые свойства изображений - цвета, текстуры и композиция - а не их словесные описания. Можно комбинировать запросы по содержанию с поиском по тексту и ключевым словам, что дает мощные возможности по нахождению данных изображений и объектов мультимедиа.

У каждого сервера поиска изображений есть каталог данных с одной или несколькими базами данных поиска изображений, где хранятся каталоги поиска изображений. В каталоге поиска изображений хранятся данные о видимых характеристиках изображений из собрания. При этом сами изображения хранятся на серверах объектов в системе IBM Content Manager for Multiplatforms. Сервер поиска изображений работает в AIX и Windows.

Информацию об установке поиска изображений смотрите в книге *Планирование и установка Enterprise Information Portal*.

Конфигурирование поиска изображений

Эти инструкции применимы только после установки поиска изображений, которая выполняется автоматически, если выбрать соединитель Content Manager Версии 7.1. Настройка поиска изображений включает следующие этапы:

1. Настройка среды
2. Конфигурирование сервера поиска изображений
3. Конфигурирование клиента поиска изображений
4. Загрузка примера изображения

Если вы используете мастер по установке в AIX: Не нужно запускать сценарии конфигурации и настройки или вводить команду конфигурирования сервера. Мастер выполнит эти задачи сам.

При установке в Windows: Вы должны выполнить эти действия.

Настройка среды

Действия настройки среды, описанные в этом разделе, надо выполнить и на сервере, и на клиенте. Для сервера поиска изображений надо задать переменные среды:

QBICTOP
Для разрешения имен файлов при конфигурировании поиска изображений

QbicImagePath
Для разрешения имен файлов изображений сервера

QbicMaskPath
Для разрешения имен файлов масок сервера

QbicSketchPath
Для разрешения имен файлов эскизов сервера

QbicTextPath
Для разрешения имен текстовых файлов сервера

На клиенте поиска изображений требуется только переменная среды QBICTOP.

Пример для AIX: В AIX запустите сценарий конфигурирования, который создает сценарий настройки, затем запустите полученный сценарий, чтобы настроить среду.

1. Запустите следующий сценарий конфигурирования:

   ```
   /usr/lpp/cmb/bin/frnconfg.iss QBICTOP
   ```

   где QBICTOP - путь к управляющим файлам (*.ini). Для QBICTOP задайте значение /user1/cmb/qbic, где /user1 - домашний каталог ID пользователя администратора поиска изображений. ID пользователя поиска изображений должен иметь права чтения/записи для этого каталога.

   Этот сценарий создаст сценарий настройки: frnsetup.iss.
2. Из домашнего каталога ID пользователя поиска изображений запустите:

   ```
   . ./frnsetup.iss
   ```

   Этот сценарий задает переменные среды для серверов и клиентов поиска изображений.

Пример для Windows: Чтобы задать переменные среды:

1. Выберите **Пуск ⟶ Настройка ⟶ Панель управления**.
2. Дважды щелкните по значку **Система**.
3. Щелкните по вкладке Среда.
4. Задайте переменные и значения, указанные в Табл. 8, введя их в соответствующих полях и нажав кнопку **Задать**.

   Требование: Для клиента поиска изображений требуется только переменная QBICTOP. Для среды клиента задайте только переменную QBICTOP.

Таблица 8. Переменные среды поиска изображений

Переменная	Значение
QBICTOP	d:\cmbroot\iss
QbicImagePath	d:\cmbroot\iss
QbicMaskPath	d:\cmbroot\iss
QbicSketchPath	d:\cmbroot\iss
QbicTextPath	d:\cmbroot\iss
Где d: - диск, на котором установлен поиск изображений.	

Конфигурирование сервера поиска изображений

Перед запуском сервера поиска изображений его надо сконфигурировать. Конфигурирование сервера состоит из начального конфигурирования и проверки соединения.

Чтобы сконфигурировать сервер:

1. Запустите интерпретатор команд, введя: qbicadm
2. Введите команду **config server**. Например,

   ```
   config server LIBSRVRN FRNADMIN PASSWORD 9999
   ```

   где LIBSRVRN - имя библиотечного сервера, FRNADMIN - ID пользователя Content Manager, PASSWORD - пароль Content Manager, а 9999 - номер порта сервера поиска изображений.

Дополнительную информацию смотрите в разделе “Проверка соединения” на стр. 114.

Конфигурирование клиента поиска изображений

Прежде, чем запускать клиент поиска изображений, в том числе программы управления системой поиска изображений, его надо сконфигурировать. Для управления системой Content Manager вы должны назначить алиас. Проверьте вашу конфигурацию, протестировав соединение.

Задание алиаса: Прежде чем использовать программу управления системой Content Manager поиска изображений, которая действует как клиент поиска изображений, надо задать хотя бы один алиас сервера.

Чтобы задать алиас:

1. Запустите интерпретатор команд, введя: `qbicadm`
2. Введите команду **add alias**. Например,

   ```
   add alias QBICSRV HOSTNAME 9999
   ```

   где `QBICSRV` - алиас, `HOSTNAME` - имя хоста сервера поиска изображений, а `9999` - номер порта сервера поиска изображений.

Проверка соединения:

Внимание:

1. Для соединения с сервером поиска изображений должен быть запущен библиотечный сервер.
2. Программе управления системой поиска изображений нужен существующий ID пользователя Content Manager. Чтобы соединение с библиотечным сервером было успешным, ID пользователя поиска изображений должен совпадать с ID пользователя библиотечного сервера. Значение этого ID пользователя по умолчанию - `frnadmin`. Если вы изменяете это значение, убедитесь, что оба ID совпадают.

Чтобы проверить соединение:

1. После конфигурирования сервера поиска изображений и задания алиаса запустите сервер, введя в командной строке сервера `commsrv`.
2. Чтобы запустить интерпретатор команд, введите: `qbicadm`.
3. В интерпретаторе команд введите команду `connect`.

   ```
   connect QBICSRV FRNADMIN PASSWORD
   ```

   где `QBICSRV` - алиас, `FRNADMIN` - ID пользователя Content Manager, `PASSWORD` - пароль Content Manager.

   После успешного соединения появится сообщение `Библиотечный сервер - LIBSRVRN`.
4. Чтобы отсоединиться от сервера, введите: `disconnect`.
5. Чтобы выйти из интерпретатора команд, введите: `quit`.

Загрузка и индексирование примеров данных

В этом разделе объясняется, как загрузить и проиндексировать примеры данных текста и изображений, используемые с примерами прикладных программ. Информация из этого раздела применима, только если вы установили соединитель Content Manager Версии 7.1 и включили опцию текстового поиска.

На компакт-диске Enterprise Information Portal поставляется несколько программ загрузки примеров. В этом разделе описана загрузка текстовых данных и

изображений при помощи программы загрузки примеров LoadSampleTSQBICDL. Можно загрузить текстовые данные и изображения по отдельности, чтобы убедиться, что обе эти возможности работают правильно.

Действия перед загрузкой данных

Перед тем, как запускать загрузчик:

1. Зарегистрируйтесь в клиенте администратора EIP. Выберите Пуск —► Программы —► **Enterprise Information Portal** for Multiplatforms 8.2 —► Управление.
2. Выберите базу данных и зарегистрируйтесь под нужными ID пользователя и паролем. Если вы выбрали базу данных по умолчанию icmnlsdb, введите в качестве ID пользователя **icmadmin**, а в качестве пароля - password. Если вы выбрали другую базу данных, введите соответствующий ID пользователя.
3. Создайте конфигурацию библиотечного сервера при помощи инструмента системного администратора Content Manager. Если для этого вам потребуется помощь, обратитесь к электронной справке программы системного администратора.
4. Измените свойства **Доступа** конфигурации библиотечного сервера:
   a. Дважды щелкните по новой конфигурации и выберите **Свойства**, чтобы открыть записную книжку Свойства.
   b. Щелкните по вкладке Доступ.
   c. Выберите радиокнопку **Неограниченное число сеансов с любой рабочей станции**.

Создание индекса текстового поиска

Перед загрузкой данных надо создать пустой индекс текстового поиска, который будет использоваться при индексировании текстовых примеров. **Совет:** Индекс текстового поиска можно создать только на сервере Content Manager Версии 6.1 или 7.1.

Чтобы создать индекс текстового поиска:

1. Запустите сервер текстового поиска на рабочей станции, где он установлен, введя команду:

   ```
   imlss -start экземпляр
   ```

   где экземпляр - имя экземпляра сервера текстового поиска, выбранное при установке или заданное при помощи командной утилиты imlcfgsv.
2. Запустите программу администратора системы Content Manager и зарегистрируйтесь в ней.
3. Выберите из выпадающего списка на верхней левой панели **Текстовый поиск**.
4. Дважды щелкните мышью по папке **Серверы поиска** на левой панели.
5. Дважды щелкните по серверу **TM**. TM - это алиас сервера поиска для сервера текстового поиска.
6. Дважды щелкните по папке **Индексы** на левой панели. Если появится сообщение RC_EMPTY_LIST, выберите в полосе меню, **Выбранные —► Новый**, чтобы создать новый индекс.
7. В окне Новый индекс определите новый индекс. Нажмите кнопку **Справка**, чтобы посмотреть подробное описание каждого поля.

   Например:

   Для Windows:

   Имя TMINDEX

Тип Точный

Файлы индекса
x:\cmbroot\ts\index\tmindex, где x - диск установки; если этот путь не существует, он создается автоматически.

Рабочие файлы индекса
x:\cmbroot\ts\work\tmindex, где x - диск установки; если этот путь не существует, он создается автоматически.

Ввод информации
Имя библиотечного сервера Content Manager.

Не изменяйте имена DLL по умолчанию для клиента и сервера.

Для AIX:

Имя TMINDEX

Тип Точный

Файлы индекса
/home/cltadmin/tsindex/index/tmindex; если этот путь не существует, он создается автоматически.

Рабочие файлы индекса
/home/cltadmin/tsindex/work/tmindex; если этот путь не существует, он создается автоматически. Пользователь должен иметь право записи в этот каталог.

Ввод информации
Имя библиотечного сервера Content Manager.

8. Нажмите кнопку **OK**.
9. Дважды щелкните по значку **TMINDEX**, чтобы открыть записную книжку Управление TMINDEX.

Создание базы данных поиска изображений, каталога и характеристик

Когда индекс текстового поиска для данных примера создан, надо создать базу данных поиска изображений и каталог для примеров изображений.

Чтобы создать базу данных поиска изображений, каталог и характеристики:

1. Запустите сервер поиска изображений на рабочей станции, где он установлен, введя команду:
   commsrv
2. Запустите программу администратора системы Content Manager и зарегистрируйтесь в ней.
3. Выберите из списка на верхней левой панели **Поиск изображений**.
4. Щелкните по папке **Серверы поиска изображений** на левой панели.
5. Щелкните по значку **QBICSRV**.
   где QBICSRV - имя сервера поиска изображений, которое вы задали при установке.
6. Щелкните правой кнопкой по папке **Базы данных** на левой панели и выберите **Новая база данных**.
7. В окне Новая база данных введите SAMPLEDB в поле **Имя** и нажмите кнопку **OK**.
8. На левой панели щелкните по папке **Базы данных**, чтобы на панели появился значок **SAMPLEDB**.

9. Щелкните по значку **SAMPLEDB**.
10. На левой панели щелкните правой кнопкой по папке **Каталоги** и выберите **Новый каталог**.
11. В окне Новый каталог введите SAMPLECAT в поле **Имя** и нажмите кнопку **OK**.
12. На левой панели щелкните по папке **Каталоги**, чтобы появился значок **SAMPLECAT**.
13. Щелкните по значку **SAMPLECAT**.
14. На левой панели щелкните правой кнопкой мыши по значку **Характеристики** и выберите **Новые характеристики**.
15. В окне Новые характеристики выберите каждую из характеристик в поле **Имя** и нажмите кнопку **Применить**. Когда все четыре характеристики будут выбраны, появится сообщение:

    ```
    Все возможные характеристики уже добавлены к каталогу.
    ```
16. Нажмите кнопку **OK**.
17. Нажмите кнопку **Отмена**.

Запуск программы загрузки

Для проверки текстового поиска и поиска изображений можно загрузить примеры данных.

Примеры изображений находятся в следующих файлах:

В Windows:

```
x:\cmbroot\samples\java\dl\samples.jar
```

В AIX:

```
/usr/lpp/cmb/samples/java/dl/samples.jar
```

Программа загрузки примеров загружает эти данные в Content Manager и индексирует их. Прочтите пролог в исходном тексте программы, где приводятся указания по вызову программ. Программы загрузки примеров:

В Windows:

```
x:\cmbroot\samples\java\dl\LoadSampleTSQBICDL.jar
```

В AIX:

```
/usr/lpp/cmb/samples/java/dl/LoadSampleTSQBICDL.jar
```

Для запуска программы загрузки примеров данных:

1. Распакуйте файлы .jar, введя:

   ```
   jar -xvf samples.jar
   ```

   Файлы будут распакованы в правильные каталоги.
2. Задайте переменные среды вашей рабочей станции, чтобы скомпилировать программу загрузки примеров; для этого:

   В Windows:

   a. Откройте в текстовом редакторе файл x:\cmbroot\cmbenv71.bat и измените первые три строки, задающие переменные среды рабочей станции:

      ```
      set CMBROOT = e:\cmbroot
      set DB2HOME = e:\sqllib
      set JAVAHOME = d:\jdk117
      ```

   b. Сохраните файл cmbenv71.bat и задайте значения переменных среды, введя:

```
cmbenv71
```

В AIX:

a. Перейдите в каталог `/usr/lpp/cmb/bin/` и задайте переменные, введя
```
. ./cmbenv71.sh
```
b. Проверьте, что подкаталоги в `/usr/lpp/cmb/samples/java/dl` и файлы примеров доступны на запись для всех пользователей.

3. Скомпилируйте программу загрузчика, введя с соблюдением регистра команду:
```
javac  LoadSampleTSQBICDL.java
```
4. **Требование:** Перед запуском программы загрузки должны быть запущены следующие серверы:
   - Библиотечный сервер
   - Сервер объектов
   - Сервер текстового поиска
   - Сервер поиска изображений

   Если вы работаете с локализованной версией Content Manager, перед запуском программы загрузки задайте для переменной `FRNDEFLANG` значение `ENU`. В AIX для задания этой переменной среды введите команду: `export FRNDEFLANG=ENU`
5. Загрузите данные примеров при помощи программы загрузки:
```
java LoadSampleTSQBICDL sampleQBIC.dat load.log frnadmin password LIBSRVRN
```
   где `frnadmin` - ID пользователя, `password` - пароль, а `LIBSRVRN` - библиотечный сервер.
6. Проверьте файл `load.log`, чтобы убедиться, что данные примера успешно загружены.

После загрузки примеров данных при помощи программы администратора системы Content Manager или средств текстового поиска командной строки проиндексируйте примеры текстовых данных.

Индексирование примеров текстовых данных

Чтобы проиндексировать примеры текстовых данных:

1. Запустите программу администратора системы Content Manager и зарегистрируйтесь в ней.
2. Выберите из выпадающего списка на верхней левой панели **Текстовый поиск**.
3. Дважды щелкните мышью по папке **Серверы поиска**.
4. Дважды щелкните по серверу **TM**. TM - это алиас сервера поиска для сервера текстового поиска.
5. Щелкните правой кнопкой по значку **нового текстового индекса** и выберите **Свойства**.
6. На странице Явные записной книжки Свойства нажмите кнопку **Обновить**.
7. В поле **Счетчик индекса** будет показано число загруженных программой загрузки документов.
8. Нажмите кнопку **Индексировать**, чтобы проиндексировать файлы.
9. Немного подождите и нажмите кнопку **Обновить**, чтобы увидеть в поле **Первичный индекс документа**, сколько документов проиндексировано.

После индексирования данных можно использовать пример прикладной программы Java для запросов к собранию или запускать простые запросы при помощи утилиты командной строки `imlsrch`.

Глава 9. Форматы документов

Форматы документов исследования информации

В этом приложении описаны форматы документов, поддерживаемые при исследовании информации.

Текстовые процессоры: Общие форматы

Текст ANSI (7- и 8-битный)	Все версии
Текст ASCII (доступны 7- и 8-битные версии)	Все версии
HTML	Версии до 3.0 (с некоторыми ограничениями)
IBM FFT	Все версии
IBM Revisable Form Text	Все версии
Microsoft Rich Text Format (RTF)	Все версии
Текст Unicode	Все версии

Текстовые процессоры: DOS

DEC WPS Plus (DX)	Версии до 4.0
DEC WPS Plus (WPL)	Версии до 4.1
DisplayWrite 2 & 3 (TXT)	Все версии
DisplayWrite 4 & 5	Версии до выпуска 2.0
Enable	Версии 3.0, 4.0 и 4.5
First Choice	Версии до 3.0
Framework	Версия 3.0
IBM Writing Assistant	Версия 1.01
Lotus Manuscript	Версии до 2.0
MASS11	Версии до 8.0
Microsoft Word	Версии до 6.0
Microsoft Works	Версии до 2.0
MultiMate	Версии до 4.0
Navy DIF	Все версии
Nota Bene	Версия 3.0
Office Writer	Версии с 4.0 по 6.0
PC-File Letter	Версии до 5.0
PC-File+ Letter	Версии до 3.0
PFS:Write	Версии A, B и C

Professional Write	Версии до 2.1
Q&A	Версия 2.0
Samna Word	Версии до 4.0
SmartWare II	Версия 1.02
Sprint	Версия 1.0
Total Word	Версия 1.2
Volkswriter 3 & 4	Версии до 1.0
Wang PC (IWP)	Версии до 2.6
WordMARC	Версии до Composer Plus
WordPerfect	Версии до 6.1
WordStar	Версии до 7.0
WordStar 2000	Версии до 3.0
XyWrite	Версии до III Plus

Текстовые процессоры: Международные

JustSystems Ichitaro	Версии 5.0, 6.0, 8.0, 9.0 и 10.0

Текстовые процессоры: Windows

AMI/AMI Professional	Версии до 3.1
Corel WordPerfect for Windows	Версии до 9.0
JustWrite	Версии до 3.0
Legacy	Версии до 1.1
Lotus WordPro (платформы Win32 / Intel)	SmartSuite 96, 97 и Millennium
Lotus WordPro (платформы Unix - только текст)	SmartSuite 97 и Millennium
Microsoft Windows Works	Версии до 4.0
Microsoft Windows Write	Версии до 3.0
Microsoft Word 97	Word 97
Microsoft Word 2000	Word 2000
Microsoft Word for Windows	Версии до 7.0
Microsoft WordPad	Все версии
Novell Perfect Works	Версия 2.0
Novell WordPerfect for Windows	Версии до 7.0
Professional Write Plus	Версия 1.0
Q&A Write for Windows	Версии 3.0
WordStar for Windows	Версия 1.0

Текстовые процессоры: Macintosh

Microsoft Word	Версии с 4.0 по 6.0
Microsoft Word 98	Word 98
WordPerfect	Версии с 1.02 по 3.0
Microsoft Works	Версии до 2.0
MacWrite II	Версия 1.1

Форматы электронных таблиц

VP Planner 3D	Версия 1.0
Enable	Версии 3.0, 4.0 и 4.5
First Choice	Версии до 3.0
Framework	Версия 3.0
Lotus 1-2-3 (DOS & Windows)	Версии до 5.0
Lotus 1-2-3 for SmartSuite	SmartSuite 97 и Millennium
Lotus 1-2-3 Charts (DOS & Windows)	Версии до 5.0
Lotus 1-2-3 (OS/2)	Версии до 2.0
Lotus 1-2-3 Charts (OS/2)	Версии до 2.0 Plus
Lotus Symphony	Версии 1.0, 1.1 и 2.0
Microsoft Excel 97	Excel 97
Microsoft Excel 2000	Excel 2000
Microsoft Excel Macintosh	Версии с 3.0 по 4.0, 98
Microsoft Excel Windows	Версии с 2.2 до 7.0
Microsoft Excel Charts	Версии с 2.x до 7.0
Microsoft Multiplan	Версия 4.0
Microsoft Windows Works	Версии до 4.0
Microsoft Works (DOS)	Версии до 2.0
Microsoft Works (Mac)	Версии до 2.0
Mosaic Twin	Версия 2.5
Novell Perfect Works	Версия 2.0
QuattroPro for DOS	Версии до 5.0
QuattroPro for Windows	Версии до 9.0
PFS:Professional Plan	Версия 1.0
SuperCalc 5	Версия 4.0
SmartWare II	Версия 1.02

Форматы баз данных

SmartWare II	Версия 1.02

Access	Версии до 2.0
dBase	Версии до 5.0
DataEase	Версия 4.x
dBXL	Версия 1.3
Enable	Версии 3.0, 4.0 и 4.5
First Choice	Версии до 3.0
FoxBase	Версия 2.1
Framework	Версия 3.0
Microsoft Windows Works	Версии до 4.0
Microsoft Works (DOS)	Версии до 2.0
Microsoft Works (Mac)	Версии до 2.0
Paradox (DOS)	Версии до 4.0
Paradox (Windows)	Версии до 1.0
Personal R:BASE	Версия 1.0
R:BASE 5000	Версии до 3.1
R:BASE System V	Версия 1.0
Q & A	Версии до 2.0
Reflex	Версия 2.0

Стандартные графические форматы

PNG - Portable Network Graphics Internet Format	Версия 1.0
Binary Group 3 Fax	Все версии
BMP (в том число RLE, ICO, CUR & os/2 DIB)	Windows
CDR (со вложенными изображениями TIFF)	Coral Draw версии 2.0 – 9.0
CGM - Computer Graphics Metafile	ANSI, CALS, NIST, Версия 3.0
CMX - формат Corel Clip Art	Версии с 5 по 6
DCX (многостраничный PCX)	Microsoft Fax
DRW - Micrografx Designer	Версия 3.1
DRW - Micrografx Draw	Версии до 4.0
DXF (двоичный и ASCII) - формат AutoCAD Drawing Interchange	Версии до 14
EMF	Windows Enhanced Metafile
EPS Encapsulated PostScript	Со вложенными изображениями TIFF
FMV - FrameMaker graphics	Векторный и растровый формат до Версии 5.0
FPX - Kodak Flash Pix	Нет определенного формата

GDF - IBM Graphics Data Format
Версия 1.0

GEM - Graphics Environment Manager Metafile
Растровый и векторный

GIF - Graphics Interchange Format
Compuserve

GP4 - формат Group 4 CALS Type I и Type II

HPGL - Hewlett Packard Graphics Language
Версия 2.0

IMG - GEM Paint Нет определенной версии

JFIF (JPEG не в формате TIFF)
Все версии

JPEG - формат Joint Photographic Experts Group
Все версии

MET - OS/2 PM Metafile Версия 3.0

PBM - Portable Bitmap Нет определенной версии

Kodak Photo CD Версия 1.0

PCD - PCX Bitmap PC Paintbrush

Perfect Works (Draw) Novell версия 2.0

PGM - Portable Graymap Нет определенной версии

PIC - Lotus 1-2-3 Picture File Format
Нет определенной версии

PICT1 & PICT2 (растровый) Macintosh Standard

PIF - IBM Picture Interchange Format
Версия 1.0

PNTG MacPaint

PPM - Portable Pixmap Нет определенной версии

Progressive JPEG Нет определенной версии

PSP - Paintshop Pro (только Win32)
Версии 5.0, 5.0.1

RND - AutoShade Rendering File Format
Версия 2.0

SDW Ami Draw Snapshot (Lotus)
Все версии

SRS - формат Sun Raster File Нет определенной версии

Targa Truevision

TIFF Версии до 6

TIFF CCITT Group 3 & 4 Fax Systems

VISO (режим Page Preview только для Версии 4) Visio 4, 5, 2000
Visio 4, 5, 2000

WMF Windows Metafile

WordPerfect Graphics [WPG и WPG2]	Версии до 2.0
XBM - X-Windows Bitmap	x10-совместимый
XPM - X-Windows Pixmap	x10-совместимый
XWD - X-Windows Dump	x10-совместимый

Форматы профессиональных графических систем

PSD - формат файлов Adobe Photoshop	Версия 4.0
AI - формат файлов Adobe Illustrator	Версии до 7.0
CDR - Corel Draw	Версии до 8.0
DSF - Micrografx Designer	Windows 95, версия 6.0
DWG - собственный формат рисунков AutoCAD	Версии с 12 по 14
IGES - Initial Graphics Exchange Specification	Версия 5.1
PDF - Portable Document Format	Acrobat версии 2.1, 3.0, 4.0, в том числе японская версия PDF
PS - Postscript	Level 2

Форматы презентаций

Microsoft PowerPoint for Macintosh	Версия 4.0, 98
Corel Presentations	Версия 8.0 и 9.0
Novell Presentations	Версии 3.0 и 7.0
Harvard Graphics for DOS	Версии 2.x и 3.x
Harvard Graphics	Версии для Windows
Freelance 96	Freelance 96
Freelance for Windows 95	SmartSuite 97 и Millennium
Freelance for Windows	Версии 1.0 и 2.0
Freelance for OS/2	Версии до 2.0
Microsoft PowerPoint for Windows	Версии до 7.0
Microsoft PowerPoint 97	PowerPoint 97
Microsoft PowerPoint 2000	PowerPoint 2000

Сжатые и кодированные форматы

ZIP PKWARE	Версии до 2.0g
GZIP	Нет определенной версии
LZA Self Extracting Compress	Нет определенной версии

LZH Compress	Нет определенной версии
Microsoft Binder	Версия 7.0, Binder 97
MIME (текстовая почта)	Нет определенной версии
UUEncode	Нет определенной версии
UNIX Compress	Нет определенной версии
UNIX TAR	Нет определенной версии

Другие

vCard Electronic Business Card	Версия 2.1
Исполняемые файлы (EXE, DLL)	Нет определенной версии
Исполняемые файлы для Windows NT	Нет определенной версии
MSG (только текст)	Формат почты Microsoft Outlook
Microsoft Project (только текст)	Project 98

Глава 10. Управление правами доступа

В этой главе описаны основные понятия и возможности управления правами доступа EIP. В ней объясняются методы, используемые для защиты вашей интеллектуальной собственности.

Защита интеллектуальной собственности

Мультимедийные объекты, хранящиеся в цифровом виде, являются интеллектуальной собственностью. Защита этих объектов может оказаться жизненно важной для вашего бизнеса, в особенности если эти объекты находятся в WWW, откуда их достаточно легко скопировать. Можно использовать методы маркировки, предоставляемые вместе с Content Manager, для предотвращения неавторизованного использования вашей интеллектуальной собственности путем маркировки мультимедийных цифровых объектов с целью защиты или введения идентификации по отпечаткам пальцев для доступа к объектам.

Можно ставить метку на значимых объектов, чтобы:

- Идентифицировать источник для предотвращения неавторизованного копирования или повторного использования. Такая метка называется ″водяным знаком″, она, как правило, видима.
- Идентифицировать получателя содержимого для предотвращения неавторизованного копирования или повторного использования. Такая метка называется ″отпечатками пальцев″ и, как правило, невидима.
- Дать контактный адрес для получения дополнительной информации.
- Дать информацию, например, время и дату, для использования в цепочке распределения с добавлением стоимости.

Цифровые объекты можно пометить до того, как объект будет доставлен заказчикам. И водяные знаки, и ″отпечатки пальцев″ могут быть наложены до доставки. Однако наложение ″отпечатков пальцев″ перед доставкой подразумевает, что получатель вам известен и что ему может понадобиться метка для динамического применения в процессе доставки. Наложение метки перед доставкой из вашей собственной управляемой среды повышает безопасность, поскольку уменьшает риск искажения информации.

Накладывать метки можно на нескольких этапах процесса управления и доставки, в зависимости от вашей ситуации. Метки можно накладывать на следующих этапах:

- Перед сохранением объекта

  Если для объекта используется общая метка (например, видимый водяной знак для идентификации владельца), можно применить эту метку до или во время сохранения объекта. В системе Content Manager можно сохранить и первоначальный, непомеченный объект, и помеченный объект. Другой вариант - сохранить только помеченный объект, а непомеченный объект держать в отдельном репозитории.
- После сохранения объекта

  Если надо пометить объекты, хранящиеся в системе Content Manager, вы можете получить объект, пометить его и либо заменить непомеченный объект помеченной версией, либо сохранить помеченный объект как отдельный элемент.
- После получения объекта

Если метка, которую надо наложить, зависит от получателя, это можно сделать динамически после получения объекта. Затем вместо первоначального пользователю можно направить помеченный объект.

Если в системе много непомеченных старых объектов, а вы не хотите тратить время или ресурсы, чтобы возвращаться и помечать каждый такой объект, можно помечать объекты по мере их получения.

Использование методов маркировки

Существует множество способов маркировки содержимого. Разные способы подходят для решения разных проблем и различаются по степени стойкости к удалению и изменениям.

Маркировки характеризуются по:

- Содержащейся информации

  Водяной знак
  Идентифицирует источник содержимого. Может содержать информацию, например, о владельце и версии объекта.

  "**Отпечатки пальцев**"
  Идентифицируют получателя содержимого. Могут содержать информацию, например, о том, куда и кому отправлен объект.

- Видимости

  Видимая
  Метка видима, ее можно заметить.

  Невидимая
  Метка скрыта в изображении.

- Целостность

  Нестойкая
  Метка портится при любом изменении.

  Стойкая
  Метка не изменяется при изменении помеченного объекта, например, при изменении размера, сжатии, повороте и усечении.

- Времени применения
  - При записи объекта
  - При сохранении объекта
  - При получении объекта для направления заказчику
  - При получении объекта на рабочей станции получателя
- Местонахождению
  - Если видимая метка используется для предотвращения незаконного повторного использования, ее можно наложить на большую часть изображения. Чтобы метку было труднее удалить, ее можно поместить на более текстурированную область изображения.
  - Если используется невидимая метка, в текстурированную область изображения данные можно встроить с наименьшими искажениями изображения.
  - Если видимая метка используется для обозначения владельца, ее можно ненавязчиво поместить в углу изображения.
  - Если используются и видимая, и невидимая метки, видимая метка должна применяться первой.
- Формат

Двоичный

Метка может представлять из себя произвольную последовательность битов, повторяющихся на всем изображении. Эта произвольная последовательность является ключом для применения метки к изображению или ее снятию.

Метка тоже может быть изображением.

Структурированные данные

В качестве метки могут использоваться встроенные текстовые данные.

Видимая маркировка

Видимая метка - это прозрачная маска, размещаемая на изображении таким образом, что видны и маска, и изображение. Видимая метка, которую трудно удалить - эффективное средство против неправомерного использования ваших объектов.

Используйте видимую метку в следующих ситуациях:

- Если хотите предоставить заказчикам возможность повторного просмотра изображения, без возможности повторного использования просмотренных копий
- Если хотите использовать изображение в качестве рекламы в WWW

Невидимая маркировка

Невидимая метка - это данные, спрятанные в изображении так, что оно остается неизмененным. Чтобы наложить, обнаружить и раскодировать такую метку, нужна специальная программа.

Используйте невидимую метку в следующих ситуациях:

- Если хотите встроить информацию, чтобы идентифицировать владельца и не допустить создания нелегальных копий помеченных объектов (нанесение водяных знаков)
- Если хотите встроить информацию для отслеживания пути распределения (метод ″отпечатков пальцев″)
- Если хотите встроить в изображение аннотацию или заголовок

Глава 11. Доступность

Этот продукт включает ряд возможностей, делающих его доступнее для лиц с физическими ограничениями. В их число входят:

- Возможность использовать клавиатуру вместо мыши для работы с любыми функциями.
- Поддержка улучшенных свойств дисплея
- Совместимость с дружелюбными технологиями
- Совместимости с возможностями доступности операционной системы
- Доступные форматы документации

Ввод и перемещение без помощи мыши

Для ввода данных и перемещения при помощи клавиатуры доступны следующие возможности:

Ввод с клавиатуры

Для работы с продуктом вместо мыши можно использовать клавиатуру.

Для пунктов меню и элементов управления есть клавиши доступа, позволяющие вам активировать элемент управления или выбирать пункт меню с помощью клавиатуры. Соответствующие этим клавишам буквы выделяются в названиях элементов управления и пунктов меню подчеркиванием.

Фокус ввода с клавиатуры

В системах на основе Windows фокус ввода с клавиатуры выделяется на экране; тем самым указывается активная область окна, в которую будут вводиться символы при нажатии клавиш.

Настройка времени ответа

В системах на основе Windows можно настраивать время ответа при помощи панели управления.

Средства облегчения работы с экраном

В клиентах есть ряд средств, улучшающих пользовательский интерфейс и облегчающих работу для пользователей со слабым зрением. К ним относятся поддержка параметров высококонтрастного вывода на экран и настраиваемых свойств шрифтов.

Высококонтрастный режим

Клиенты поддерживают высококонтрастный режим, задаваемый в операционной системе. Эта возможность поддерживает высокий контраст между цветами текста и фона.

Параметры шрифтов

В системах на основе Windows вы можете выбрать цвет, размер и тип шрифта, используемого в меню и для диалоговых окон. Клиент позволяет вам также выбрать шрифт для списка документов.

Независимость от цвета

Чтобы использовать любые функции этого продукта, пользователям не требуется различать цвета.

Совместимость с технологиями для людей с физическими недостатками

Клиенты совместимы с программами чтения с экрана (такими, как Narrator и Via Voice). Клиенты обладают необходимыми свойствами, позволяющими таким программам облегчения работы получать информацию с экрана и делать ее доступной для слепых пользователей.

Удобный формат документации

Документация для этого продукта доступна в формате PDF. Эти файлы PDF можно преобразовать в HTML или текстовые файлы при помощи бесплатных программ фирмы Adobe на сайте access.adobe.com. Это позволяет пользователям просматривать документацию, используя предпочтения экрана, заданные для их браузеров. Это позволяет также использовать программы чтения с экрана и другие технологии для людей с физическими недостатками.

Замечания

Эта публикация разрабатывалась для продуктов и услуг, предлагаемых в США.

IBM может не предоставлять продукты, услуги или средства, описываемые в этом документе, в других странах. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному торговому представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права фирмы IBM на интеллектуальную собственность. Однако при этом пользователь сам несет ответственность за оценку и проверку работы с другими (не IBM) продуктами, программами и услугами.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данной публикации. Получение этого документа не означает предоставления каких-либо лицензий на эти патенты. Запросы относительно лицензий направляйте по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106, Japan

Следующий абзац неприменим в Великобритании или в любой другой стране, где подобные оговорки противоречат местному законодательству: INTERNATIONAL BUSINESS MACHINES CORPORATION ПРЕДОСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ “КАК ЕСТЬ”, БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ (НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ) ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ СОБЛЮДЕНИЯ АВТОРСКИХ ПРАВ, РЫНОЧНОЙ ПРИГОДНОСТИ ИЛИ СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ. В некоторых странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

В данной публикации могут встретиться технические неточности или типографские опечатки. В публикацию время от времени вносятся изменения, которые будут отражены в ее последующих изданиях. IBM оставляет за собой право в любое время вносить усовершенствования и/или изменения в описанные в этом замечании продукты и/или программы.

Ссылки на Web-сайты не-IBM приводятся только для вашего удобства и ни в коей мере не должны рассматриваться как рекомендации пользоваться этими

Web-сайтами. Материалы на этих Web-сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими Web-сайтами несете вы сами.

IBM может использовать или распространять информацию так, как сочтет нужным, без каких-либо обязательств с ее стороны.

Если обладателю лицензии на данную программу понадобятся сведения о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

IBM Corporation
J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003
U.S.A.

Такая информация может быть предоставлена на определенных условиях (в некоторых случаях к таким условиям может относиться оплата).

Лицензированная программа, описанная в данном документе, и все лицензированные материалы, доступные вместе с ней, предоставляются IBM на условиях Пользовательского соглашения IBM.

Все приводимые здесь данные о производительности были получены в контролируемой среде. Результаты, полученные в других средах, могут значительно отличаться от них. Часть измерений могла проводиться в системах на уровне разработки, и нет никаких гарантий, что на обычных компьютерах будут получены те же результаты. Более того, некоторые результаты могли быть получены путем экстраполяции. Реальные результаты могут быть другими. Пользователи должны проверить данные в своей собственной среде.

Информация о продуктах других фирм была получена от поставщиков этих продуктов, из их опубликованных объявлений или из других общедоступных источников. IBM не проверяла эти продукты и не может подтвердить точность характеристик, совместимость или иные заявления, связанные с продуктами других фирм. Вопросы, касающиеся характеристик продуктов других фирм (не IBM) следует адресовать поставщикам этих продуктов.

Все утверждения о будущих планах и намерениях IBM могут быть изменены или отменены без уведомлений, и описывают исключительно цели фирмы.

В этой публикации содержатся примеры данных и отчетов, используемых при выполнении текущих служебных задач. Чтобы проиллюстрировать эти задачи с максимальной наглядностью, в примерах используются имена физических лиц, названия компаний, фирм и продуктов. Все эти имена и названия являются вымышленными, и всякое сходство с именами, названиями и адресами, используемыми в реальной предпринимательской деятельности, являются не более чем совпадением.

ЛИЦЕНЗИЯ НА ПРАВО КОПИРОВАНИЯ:

Эта информация содержит примеры исходных текстов прикладных программ, которые иллюстрируют приемы программирования на различных платформах. Вы можете копировать, модифицировать и распространять эти программы примеров в любой форме без платы фирме IBM в целях разработки, использования, продажи или распространения прикладных программ, соответствующих программному

интерфейсу платформы, для которой написаны примеры. Эти примеры не были тщательно протестированы при всех возможных условиях. Поэтому IBM не может гарантировать надежность, возможность обслуживания и работоспособность этих программ и не подразумевает таких гарантий. Разрешается копировать, изменять и распространять эти примеры программ в любой форме без оплаты фирме IBM для целей разработки, использования, сбыта или распространения прикладных программ, соответствующих интерфейсам прикладного программирования IBM.

Торговые марки

Следующие термины являются товарными знаками корпорации International Business Machines в Соединенных Штатах и/или других странах:

IBM
400
Advanced Peer-to-Peer Networking
AIX
AIXwindows
APPN
AS/400
C Set ++
CICS
DATABASE 2
DataJoiner
DB2
DB2 Universal Database
DisplayWrite
e-business
HotMedia
Hummingbird
ImagePlus
IMS
Micro Channel
MQSeries
MVS/ESA
NetView
OS/2
OS/390
PAL
PowerPC
PTX
QBIC
RS/6000
SecureWay
SP
VideoCharger
Visual Warehouse
VisualAge
VisualInfo
WebSphere

Approach, Domino, Lotus, Lotus 1-2-3, Lotus Notes и SmartSuite - товарные знаки или зарегистрированные товарные знаки Lotus Development Corporation в Соединенных Штатах и в других странах.

Intel и Pentium - товарные знаки или зарегистрированные товарные знаки Intel Corporation Corporation в Соединенных Штатах и в других странах.

Microsoft, Windows и Windows NT - зарегистрированные товарные знаки Корпорации Microsoft в США и/или других странах.

Java и все основанные на Java товарные знаки и логотипы - товарные знаки или зарегистрированные товарные знаки Sun Microsystems, Inc. в США и/или других странах.

UNIX - зарегистрированный товарный знак The Open Group в США и в других странах.

Названия других компаний, продуктов и услуг могут быть товарными знаками или марками сервиса других фирм.

Глоссарий

В данном глоссарии приводятся определения терминов и сокращений, используемые в этой системе. *Курсивом* выделены термины, определения которых представлены в других статьях данного глоссария.

A

ADSM. Смотрите *Tivoli Storage Manager.*

API. Смотрите *интерфейс прикладного программирования*

Audio/Video Interleaved (AVI). Спецификация файла RIFF (*Resource Interchange File Format*), позволяющая чередовать в файле аудио- и видеоданные. Отдельные дорожки можно поместить в чередующиеся порции для воспроизведения или записи при поддержании последовательного доступа к файловому устройству.

AVI. Смотрите *Audio/Video Interleaved.*

B

BLOB. Смотрите *двоичный большой объект.*

C

CGI. Смотрите *Общий интерфейс шлюза.*

CIF. Смотрите *общий файл обмена.*

CIU. Смотрите *общий блок обмена.*

Common Gateway Interface (CGI). Стандарт для обмена информацией между Web-сервером и программами, которые по отношению к нему являются внешними. Такие внешние программы могут быть написаны на любом языке программирования, поддерживаемом операционной системой, в которой работает Web-сервер. Смотрите *сценарий CGI.*

D

DDO. Смотрите *динамический объект данных.*

DTD. Смотрите *определение типа документа.*

E

Extensible Markup Language (XML). Стандартный метаязык для определения языков разметки, основанный на SGML и являющийся его подмножеством. В XML исключены наиболее сложные и редко используемые части SGML, что упрощает написание программ и обработку типов документов, работу с структурированной информацией, ее передачу и совместное использование в различных компьютерных системах. Использование XML не требует программ высокой надежности для сложной обработки данных, что необходимо для SGML. XML разработан при содействии World Wide Web Consortium (W3C).

F

H

HTML. Смотрите *язык гипертекстовой разметки.*

I

Image Object Content Architecture (IOCA). Набор структур, используемых для обмена изображениями и для их вывода.

IOCA. Смотрите *архитектура содержимого объектов изображений.*

J

JavaBeans. Не зависящая от платформы технология программных компонентов, позволяющая строить многократно используемые компоненты Java, называемые “компонентами bean”. После построения beans можно сделать доступными для использования другими разработчиками программного обеспечения или прикладными программами Java. При помощи JavaBeans разработчики программного обеспечения могут применять и компоновать beans в графической среде разработки с возможностями перетаскивания.

Joint Photographic Experts Group (JPEG). (1) Группа, разработавшая стандарт для сжатия оцифрованных естественных (с непрерывными переходами тонов) изображений. (2) Стандарт для неподвижных изображений, разработанный этой группой.

JPEG. Смотрите *Joint Photographic Experts Group.*

K

L

LAN. Смотрите *локальная сеть.*

M

Multipurpose Internet Mail Extensions (MIME) . Смотрите *тип MIME*.

N

O

OLE. Смотрите *связывание и встраивание объектов*.

P

PID. Смотрите *постоянный идентификатор.*

Q

QBIC. Смотрите *запрос по содержимому изображения.*

R

Resource Interchange File Format (RIFF) . Формат для хранения звука или графики при их воспроизведении на различных типах компьютерного оборудования.

RIFF. Смотрите *Resource Interchange File Format*.

RMI-сервер (RMI server). Сервер, обеспечивающий реализацию модели распределенных объектов *RMI* Java.

S

T

Tivoli Storage Manager (TSM). Продукт типа *клиент/сервер*, который дает возможность управлять хранением и предоставляет службы доступа к данным в гетерогенной среде. Он поддерживает различные способы взаимодействия, содержит средства управления, обеспечивающие резервное копирование и хранение файлов, а также позволяет производить планирование операций по резервному копированию.

TSM. Смотрите *Tivoli Storage Manager.*

U

V

W

X

XDO. Смотрите *расширенный объект данных.*

XML. Смотрите *Extensible Markup Language*.

A

абстрактный класс (abstract class). *Класс* объектно-ориентированного программирования, который представляет собой понятие; классы, полученные на его основе, представляют собой реализации этого понятия. Вы не можете сконструировать объект абстрактного класса; то есть, создать экземпляр абстрактного класса нельзя.

анализ информации. Автоматизированная процедура извлечения важной информации из текста (суммирование), поиска доминирующих тем в наборе документов (категоризация) и поиска нужных документов на основе мощного и гибкого механизма запросов.

атрибут (attribute). Единица данных, описывающая определенную характеристику или свойство (например, имя, адрес, возраст и т.п.) элемента; ее можно использовать для поиска этого элемента. У атрибута есть тип, которые описывает допустимые значения данных, хранящихся в этом атрибуте, и значение в допустимом диапазоне. Пример атрибута - информация о файле в мультимедийной файловой системе, такая как заголовок, время воспроизведения или тип кодирования (MPEG1, H.263 и т.п.). Для Enterprise Information Portal смотрите также *атрибут объединения* и *собственный атрибут*.

атрибут объединения (federated attribute). Категория метаданных Enterprise Information Portal, отображенная в *собственные атрибуты* на одном или нескольких *контент-серверах*. Например, атрибут объединения `номер полиса` может в Content Manager отображаться в *атрибут* `policy num`, а в Content Manager ImagePlus for OS/390 - в атрибут `policy ID`.

Б

библиотечный клиент. Компонент системы Content Manager, который поддерживает низкоуровневый программный интерфейс библиотечной системы. В библиотечный клиент входят API, составляющие часть комплекта разработчика программ.

библиотечный сервер. Компонент системы Content Manager, который хранит и обрабатывает запросы об *элементах* и управляет этими запросами.

В

выдвижение (staging). Процесс перемещения хранящегося *объекта* с неподключенного или низкоприоритетного устройства на подключенное или высокоприоритетное, обычно по требованию системы или по заказу пользователя. Когда пользовательские требования на объект сохраняются в постоянной памяти, рабочая копия объекта записывается в *кэш менеджера ресурсов* (на сцену).

вызов удаленного метода (Remote Method Invocation - RMI). Набор API, обеспечивающий распределенное программирование. Объект в одной системе Java Virtual Machine (JVM) может вызывать методы для объектов в других JVM.

высвободить (release). Отменить критерий приостановки для *элемента.* Высвобождение приостановленного элемента произойдет, если будет достигнуто соответствие критериям или если пользователь с соответствующими полномочиями перезапишет критерии и вручную высвободит элемент.

Г

Гбайт (GB). Смотрите *гигабайт*.

гигабайт (gigabyte, GB). (1) Для памяти процессора, реальной и виртуальной памяти, а также для пропускной способности канала - 2^{30} или 1073741824 бита. (2) Для объема дисковой памяти и объема передаваемой информации - 1 000 000 000 байт.

группа пользователей (user group). Группа из одного или нескольких отдельных *пользователей*, имеющая единое групповое имя.

Д

двоичный большой объект (BLOB, binary large object). Последовательность байтов, объем которой находится в диапазоне от 0 байт до 2 Гбайт. С такой строкой не связана ни кодовая страница, ни набор символов. В виде BLOB хранятся объекты изображений, аудио- и видеообъекты.

динамический объект данных (dynamic data object - DDO). В прикладных программах: общий способ представления сохраненного объекта, который позволяет перемещать этот объект в место хранения и из него.

документ (document). *Элемент*, который хранится, вызывается и передается из системы Content Manager в другую систему или пользователю как отдельная единица. Ожидается, что элемент с *семантическим типом* документ содержит информацию, которая образует документ, хотя и не обязательно реализует при этом модель документа Content Manager.

Элемент, созданный с документным типом элементов (конкретная реализация модели документа Content Manager), должен содержать части документа. Документные типы элементов можно использовать для создания элементов с семантическим типом документов или папок.

Части документа могут иметь разные типы содержимого, включая, например, текст, изображения и электронные таблицы.

дочерний компонент (child component). Дополнительный второй или низший уровень иерархического *типа элементов*. Каждый дочерний компонент непосредственно связан с вышестоящим уровнем.

З

запрос по содержимому изображения (query by image content - QBIC). Технология запроса, позволяющая искать не текст, а визуальное содержание изображения, называемое характеристиками. Используя QBIC, можно искать объекты по таким визуальным характеристикам, как цвет и текстура.

И

индексировать (index). Добавлять или редактировать значения атрибутов, идентифицирующих определенный *элемент* или *объект*, чтобы иметь возможность получать его позже.

индексный класс (index class). Смотрите *тип элемента*.

интерфейс прикладного программирования (application programming interface - API). Программный интерфейс, обеспечивающий возможность взаимодействия приложений друг с другом. API - это набор конструкций или операторов языка программирования, которые могут добавляться в код прикладной программы, чтобы обеспечить выполнение специальных функций и служб, предоставляемых базовой лицензионной программой.

итерация (iterator). Класс конструкций, который позволяет перебирать объекты в наборе по одному.

К

класс (class). В объектно-ориентированной разработке или программировании: модель или шаблон, которые можно инициировать для создания объектов с общим определением и, следовательно, с общими свойствами, операциями и режимами. Объект является экземпляром класса.

классификация типов элементов (item type classification). Категоризация в пределах *типа элементов* для дальнейшей идентификации *элементов* данного типа. Все

элементы одного типа имеют одну и ту же классификацию типов элементов.

В Content Manager задана следующая классификация типов элементов: *папка, документ*, объект, видео, изображение и текст; пользователи могут определять свои собственные классификации типов объектов.

класс соединителя (connector class). *Класс* объектно-ориентированного программирования, который обеспечивает стандартный доступ к собственным API определенных *серверов содержимого.*

клиент/сервер (client/server). Модель взаимодействия при распределенной обработке данных, при которой программа на одном узле посылает требования программе на другом узле и ждет ее ответа. Программу, посылающую требование, называют клиентом, а отвечающую программу - сервером.

ключевое поле (key field). Смотрите *атрибут.*

комбинированный поиск (combined search). Запрос, в котором сочетаются следующие типы поиска: *параметрический* поиск, текстовый поиск или поиск изображений.

компонент (component). Общий термин для *корневого компонента* и *дочернего компонента.*

конструкция (constructor). В языках программирования: метод, имя которого совпадает с именем класса и который используется для создания и инициализации объектов этого класса.

корневой компонент (root component). Первый или единственный уровень иерархического *типа элементов*, состоящий из определенных системой и определенных пользователем *атрибутов*.

критерий поиска (search criteria). В Enterprise Information Portal - конкретные поля, заданные администратором в *шаблоне поиска* для ограничения или дальнейшего определения возможностей выбора у *пользователей.*

курсор (cursor). Именованная управляющая структура, которая в прикладной программе позволяет указать определенную строку в некотором упорядоченном наборе строк. Курсор позволяет получать строки из этого набора.

кэш (cache). Буфер специального назначения, меньше и быстрее основной памяти; используется для хранения копии часто требуемых данных. Использование кэша сокращает время доступа, но может увеличить требования к памяти.

кэш менеджера ресурсов (resource manager cache). Область рабочей памяти для *менеджера ресурсов*. Другое ее название - *сцена.*

кэш менеджера ресурсов (staging area). Область рабочей памяти для *менеджера ресурсов*. Другое название - *кэш менеджера ресурсов*.

кэш сервера объектов (object server cache). Смотрите *кэш менеджера ресурсов*.

Л

локальная сеть (local area network, LAN). Сеть, в которой набор устройств соединен друг с другом для передачи информации; может быть соединена с сетью большего размера.

М

макет (overlay). Набор предопределенных данных (линий, теней, текста, рамок или логотипов), объединяемых при печати с переменными данными на странице.

менеджер папок. Модель Content Manager для управления такими данными, как электронные документы и папки. API менеджера папок можно использовать как первичный интерфейс между вашими прикладными программами и контент-серверами Content Manager.

менеджер ресурсов. Компонент системы Content Manager, который управляет *объектами*. На эти объекты ссылаются *элементы*, которые хранятся на *библиотечном сервере*.

метод (method). В Java-разработках или в Java-программировании: программный компонент, который реализует режим, заданный операцией. Синоним этого термина в C++ - функция элемента.

минимальный клиент (thin client). Клиент с малым объемом установленных программных средств или вообще без них, но имеющий доступ к программным средствам, которыми управляют и которые предоставляют соединенные с ним сетевые серверы. Минимальные клиенты представляют собой альтернативу полнофункциональным клиентам (например, рабочим станциям).

мощность (cardinality). Число строк в таблице базы данных.

мультимедиа (multimedia). Объединение различных элементов (текста, графики, звука, неподвижных изображений, видео, анимации) для воспроизведения и управления ими при помощи компьютера.

мультимедийная файловая система (multimedia file system). *Файловая система*, оптимизированная для хранения и считывания видео- и аудиофайлов.

Н

набор привилегий (privilege set). Совокупность *привилегий* для работы с компонентами и функциями

системы. Администратор дает наборы привилегий пользователям (задаваемым ID) и *группам пользователей*.

надкласс (superclass). *Класс*, производным которого является какой-то другой класс. Между классом и надклассом могут находиться один или несколько классов.

О

обмен (interchange). Возможность импорта или экспорта изображения вместе с его индексом из одной системы Content Manager ImagePlus for OS/390 в другую систему ImagePlus с использованием *общего файла обмена* или *общего блока обмена*.

обработчик пользователя (user exit). Точка в поставляемой IBM программе, в которой управление передается подпрограмме обработчика пользователя.

общий блок обмена (common interchange unit, CIU). Независимый блок передачи для общего файла обмена (CIF). Это часть CIF, определяющая отношение с принимающей базой данных. CIF может содержать несколько CIU.

общий файл обмена (common interchange file, CIF). Файл, содержащий один поток данных ImagePlus Interchange Architecture (IPIA).

объединенный поиск (federated search). Сгенерированный в Enterprise Information Portal запрос, обеспечивающий одновременный поиск данных на одном или нескольких *контент-серверах*, которые могут быть разнородными.

объединенный склад данных (federated datastore). Виртуальное представление для обозначения любого числа указанных *контент-серверов*, например, серверов Content Manager.

объединенный текстовый индекс (federated text index). Объект метаданных Enterprise Information Portal, отображенный на один или несколько *собственных текстовых индексов* на одном или нескольких *контент-серверах*.

объект (object). Любое цифровое содержимое, которое пользователь может сохранять, получать и использовать как единое целое, например, изображения *JPEG*, аудиофайлы MP3, видеофайлы *AVI* и фрагменты текста из книг.

объект объединения (federated entity). Объект метаданных Enterprise Information Portal, состоящий из *атрибутов объединения* и (необязательно) связанный с одним или несколькими *текстовыми индексами объединения*.

определение сервера (server definition). Характеристики конкретного *контент-сервера*, которые позволяют однозначно его идентифицировать в Enterprise Information Portal.

определение типа документа (document type definition, DTD). Правила, определяющие структуру для определенного класса документов XML. Определение типа документов определяет структуру с элементами, атрибутами и обозначениями и задает ограничения, как каждый элемент, атрибут и обозначение могут использоваться с определенным классом документов. DTD аналогично схеме базы данных и полностью описывает структуру определенного языка разметки.

определение типа сервера (server type definition). Список заданных администратором характеристик, которые позволяют однозначно идентифицировать настроенный сервер определенного типа в Enterprise Information Portal.

отображение пользователя (user mapping). Связь между ID пользователей и паролями Enterprise Information Portal и соответствующими ID пользователей и паролями на одном или нескольких контент-серверах. Отображения пользователей обеспечивают единую регистрацию в Enterprise Information Portal и на нескольких *контент-серверах*.

П

пакет (package). Собрание родственных *классов* и интерфейсов, обеспечивающих защиту доступа и управление пространством имен.

папка (folder). *Элемент* любого *типа элементов* (независимо от классификации), с *семантическим типом* папка. Любой элемент с семантическим типом папка содержит особые функциональные возможности папки, обеспечиваемые Content Manager, в дополнение ко всем возможностям нересурсного элемента и дополнительным возможностям классификации типа элементов (например, *документ* или ресурсный элемент). Папки могут содержать любое число элементов любого типа, в том числе документы и подпапки. Папки индексируются по *атрибутам*.

параметрический поиск (parametric search). Запрос информации об *объектах*, созданный на основе *свойств* этих объектов.

перечень сервера (server inventory). Полный список *собственных объектов* и *собственных атрибутов* на указанных *контент-серверах*.

подкласс (subclass). *Класс*, который является производным от другого класса. Между классом и подклассом могут находиться один или несколько классов.

поднабор индексного класса (index class subset). В ранних версиях Content Manager - представление

индексного класса, используемое прикладной программой для хранения, вызова и вывода папок и объектов.

подпрограмма обработчика пользователя (user exit routine). Написанная пользователем подпрограмма, которой передается управление в заранее заданных точках вызова *обработчика пользователя*.

пользователь (user). В Enterprise Information Portal - тот, кто идентифицируется программой управления Enterprise Information Portal.

постоянный идентификатор (persistent identifier - PID). Идентификатор, обеспечивающий уникальную идентификацию *объекта* независимо от того, где он хранится. PID состоит из ID элемента и его местонахождения.

построить изображение (render). Преобразовать данные, содержание которых обычно не является визуальным, для вывода в виде изображения. В Content Manager документы текстового процессора для вывода можно преобразовать в изображение.

потоковые данные (streamed data). Любые данные, пересылаемые через сетевое соединение с определенной скоростью. Поток может состоять из данных одного типа или представлять собой комбинацию типов. Скорости данных, измеряемые в битах в секунду, различны для различных типов потоков и сетей.

представление индексного класса (index class view). В ранних версиях Content Manager - термин, используемый в API для *поднаборов индексных классов*.

привилегия (privilege). Право получать доступ к указанному *объекту* указанным способом. К привилегиям относятся права на создание, удаление и выбор объектов, хранящихся в системе. Привилегии назначаются администратором.

приостановить (suspend). Удалить *объект* из *рабочего потока* и задать критерий приостановки для его последующей активации. Последующая активация объекта позволяет продолжить его обработку.

программа клиента. Программа, написанная с использованием Интернет- или объектно-ориентированных API для обращения к *контент-серверам* с системы Enterprise Information Portal.

P

рабочее состояние (work state). Состояние отдельного *рабочего элемента*, *документа* или *папки*.

рабочий пакет (work packet). В Enterprise Information Portal Версии 7.1 - собрание *документов*, направляемое из одной точки в другую. Пользователи получают доступ к рабочим пакетам и работают с ними посредством *рабочих списков.*

рабочий поток (workflow). В Enterprise Information Portal - последовательность *рабочих шагов* и правила, управляющие этими шагами, определяющие последовательность передачи *рабочего пакета*, *документа* или *папки* при обработке.

Например, `принятие страхового иска` описывает процесс действий, которые надо выполнить с отдельным страховым иском, чтобы принять его.

рабочий список (worklist). Собрание *рабочих элементов*, *документов* или *папок*, назначаемых пользователю.

рабочий элемент (work item). В рабочем потоке ранних версий Content Manager и расширенном рабочем потоке Enterprise Information Portal - рабочая операция, выполняемая в пределах *рабочего потока.*

рабочий этап (work step). Отдельная точка в *рабочем потоке* или в *процессе маршрутизации документов*, через которую должны проходить отдельные *рабочие элементы*, *документы* или *папки.*

ранг (rank). Целочисленное значение, обозначающее релевантность данной части по отношению к результатам запроса. Чем выше ранг, тем ближе соответствие.

расширенный объект данных (extended data object - XDO). В прикладных программах: общий способ представления сохраненного комплексного мультимедийного *объекта*, который позволяет перемещать этот объект в место хранения и из него. XDO обычно содержатся в *DDO*.

C

свойство (property). Характеристика *объекта*, которая описывает его. Свойство можно изменить или модифицировать. Режим ввода - пример свойства объекта.

связывание и встраивание объектов (Object Linking and Embedding -OLE). Спецификация Microsoft для связывания и встраивания программ с целью их активации из других программ.

связь (link). Направленное взаимоотношение между двумя *элементами*: исходным элементом и элементом назначения. Набор связей можно использовать для моделирования ассоциаций ″один со многими″. Сравните со *ссылкой*.

семантический тип (semantic type). Использование или правила для *элемента.* В Content Manager заданы три семантических типа: основной, комментарий и примечание; пользователи могут также определять свои собственные семантические типы.

сервер мультимедиа (media server). Компонент системы Content Manager на платформе AIX, используемый для хранения видеофайлов и доступа к ним.

сервер объектов. Смотрите *менеджер ресурсов*.

сервер содержимого (content server). Программная система, в которой хранятся мультимедийные данные и бизнес-данные, а также соответствующие имя метаданные, которые необходимы пользователям для работы с этими данными. Примеры контент-серверов: Content Manager и Content Manager ImagePlus for OS/390.

сетевой табличный файл (network table file). Текстовый файл, содержащий специфичную для системы информацию конфигурации для каждого узла системы Content Manager. У каждого узла системы должен быть свой сетевой табличный файл, где указаны узлы и списки, с которыми он должен соединяться.

Этот файл носит имя FRNOLINT.TBL.

символ подстановки (wildcard character). Специальный символ, такой как звездочка (*) или знак вопроса (?), который можно использовать для представления одного или нескольких символов. Символ подстановки может заменять любой символ или группу символов.

склад данных (datastore). (1) Общий термин для обозначения места (например, системы базы данных, файла или каталога) хранения данных. (2) В прикладной программе - виртуальное представление *контент-сервера*.

собрание (collection). Группа объектов со сходным набором правил управления, помещенная в хранилище.

собрание объединения (federated collection). Группа объектов, полученная в результате *объединенного поиска.*

собственные атрибуты (native attributes). Характеристики объекта, управление которыми осуществляется на определенном *контент-сервере* и которые присущи только этому контент-серверу. Например, на контент-сервере Content Manager собственным атрибутом может являться *ключевое поле* `policy num`, в то время как на контент-сервере Content Manager OnDemand собственным атрибутом может являться поле `policy ID`.

собственный объект (native entity). *Объект*, управление которым осуществляется на определенном *контент-сервере* и который состоит из *собственных атрибутов. Например индексные классы Content Manager - это собственные объекты, составленные из ключевых полей Content Manager.*

собственный текстовый индекс (native text index). Индекс текстовых *элементов*, управление которыми осуществляется на определенном *контент-сервере*. Например, один текстовый индекс поиска на контент-сервере Content Manager.

состояние рабочего потока (workflow state). Состояние *рабочего потока* в целом.

список действий (action list). Одобренный список действий, заданный системным администратором или другим *координатором рабочего потока*, которые пользователю разрешено выполнять в *рабочем потоке* или в процессе маршрутизации документа.

список управления доступом (access control list). Список, включающий в себя один или несколько ID пользователей или групп пользователей и присвоенных им *привилегий*. Списки управления доступом применяются для управления доступом пользователей к *шаблонам поиска* в системе Enterprise Information Portal.

ссылка (reference). Однонаправленная одиночная ассоциация между корневым или *дочерним компонентом* и другим *корневым компонентом*. Сравните со *связью*.

строка запроса (query string). Символьная строка, которая задает свойства и значения свойств для запроса. Можно создать строку запроса в приложении, а затем передать ее запросу.

сценарий CGI (CGI script). Компьютерная программа, выполняющаяся на Web-сервере и использующая стандарт *Common Gateway Interface (CGI)* для выполнения задач, которые Web-сервер обычно не выполняет (например, получение доступа к базе данных и обработка форм). Сценарий CGI представляет собой программу GGI, написанную на таком языке сценариев, как Perl.

T

Тип MIME (MIME type). Стандарт Интернета для идентификации типа объекта, передаваемого по Интернету. Типы MIME включают в себя несколько вариантов аудио-, графических и видеообъектов. У каждого объекта есть тип MIME.

тип элементов (item type). Шаблон для определения и последующего поиска *элементов*, состоящий из *корневого компонента*, нескольких возможных *дочерних компонентов* и классификации.

том TSM (TSM volume). Логическая область хранения данных, которой управляет *Tivoli Storage Manager.*

том (volume). Понятие, соответствующее реальному физическому устройству или носителю, где хранятся объекты системы.

У

унифицированный указатель ресурсов (uniform resource locator, URL). Последовательность символов, представляющая информационные ресурсы на компьютере или в сети, например, в Интернете. Эта последовательность символов включает в себя сокращенное имя протокола, используемого для доступа к информационному ресурсу, а также информацию, используемую этим протоколом для поиска информационного ресурса. Например, в Интернете

используются такие сокращенные имена протоколов доступа к различным информационным ресурсам: http, ftp, gopher, telnet и news.

управление доступом (access control). Функция, благодаря которой доступ к тем или иным функциям и сохраненным *объектам* предоставляется только авторизованным пользователям и только разрешенными способами.

устройство хранения мультимедиа (media archiver). Физическое устройство, используемое для хранения данных потокового типа (аудио и видео). Пример устройства хранения мультимедиа - VideoCharger.

Ф

файл README (README file). Файл, который необходимо просмотреть перед установкой или запуском программы, к которой прилагается этот файл. Обычно файл README содержит последнюю информацию о продукте, информацию по установке или советы по использованию программного продукта.

файловая система (file system). В AIX - способ разбиения жесткого диска на разделы для хранения данных.

формат данных (data format). Смотрите *тип MIME*.

Х

характеристика (feature). Информация о содержании изображения, которая хранится на сервере поиска изображений. Кроме того, свойство изображения, которое программы поиска изображений используют для определения соответствий. Используется четыре характеристики *QBIC* - усредненный цвет, гистограмма цветов, позиционный цвет и текстура.

хронологический журнал (history log). Файл, где хранится запись действий для *рабочего потока.*

хэндл (handle). Символьная строка, соответствующая объекту и используемая для вызова этого объекта.

Ц

Ч

часть (part). Смотрите *объект*.

Ш

шаблон поиска (search template). Форма, состоящая из *критериев поиска*, разработанных администратором для определенного типа объединенного поиска. Администратор также указывает *пользователей* и *группы пользователей*, которые могут получать доступ к данному шаблону поиска.

шлюз (gateway). Функциональное устройство, связывающее две компьютерные сети с разными архитектурами. Шлюз соединяет сети или системы с разными архитектурами. Мост соединяет сети или системы с одинаковыми или похожими архитектурами.

Э

элемент данных (item). Общий термин для обозначения наименьшей единицы информации, которой управляет сервер Enterprise Information Portal. У каждого элемента есть идентификатор. Например, элементом данных может быть *папка* или *документ*.

Я

язык гипертекстовой разметки (Hypertext Markup Language - HTML). Язык разметки, соответствующий стандарту SGML, который в первую очередь предназначен для поддержки вывода на экран текстовой или графической информации, содержащей гипертекстовые связи.

Индекс

Р

С

Т

Ф

Ш

Номер программы: 5724-B43

Напечатано в Дании

SH43-0216-01