OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.738.52

ИНТЕГРАЦИЯ КОРПОРАТИВНЫХ ДАННЫХ НА ОСНОВЕ WIKI-СИСТЕМ В УСЛОВИЯХ СЛАБОЙ СВЯЗАННОСТИ ИСТОЧНИКОВ

Галушка И.Н., Оксанич И.Г., Щербак С.С.

Кременчугский национальный университет имени Михаила Остроградского,
г. Кременчуг, 39600, Украина.

ilona.galushka@ya.ru

В работе рассмотрены технические аспекты совместного применения Wiki-систем с хранилищами триплетов RDF для организации электронного документооборота предприятий с территориально-распределенной структурой, проведен анализ существующих подходов и способов интеграции корпоративных данных и приведена соответствующая классификация. Разработаны способы интеграции корпоративных данных на основе концепции связанных данных предприятия, и идентификации объектов в корпоративных документах, как средства формирования терминологической базы предприятия.
Ключевые слова: сервисная шина предприятия; информационное пространство; связанные данные; модель предметной области; семантика; интеграция данных

Введение

Современные рыночные условия вынуждают предприятия искать новые пути для повышения своей рентабельности, например, путем внедрения высокоэффективных систем электронного документооборота. Подобные внедрения сталкиваются с необходимостью интеграции множества различных источников данных предприятия в единое информационное пространство, компоненты которого связаны между собой единым программным интерфейсом системы электронного документооборота. Зачастую источники между собой не связаны или имеют слабовыраженную связь, что приводит к необходимости решения ряда проблемных задач, нацеленных на повышение уровня связности данных в системе и интеграции существующих источников в согласованное информационное пространство связанных данных предприятия с единой точкой входа и развитыми унифицированными средствами "бесшовного" добавления новых источников.

В последнее время одним из популярных подходов к организации корпоративного электронного документооборота является использование Вики-систем (англ. Enterprise Wiki), наиболее известной представительницей которых является MediaWiki. Эта система лежит в основе всемирно известного источника информации Википедия.

Популярность Wiki-систем обусловлена прежде всего удобным веб-ориентированным многопользовательском интерфейсов для создания и редактирования как структурированной так и не структурированной информации с поддержкой версионности и многоязычности. К сожалению, автоматическая обработка информации в Wiki - системах затруднена из-за отсутствия развитых средств для обеспечения машинного понимания и чтения (англ. Machine readable). Одной из проблем на этом пути является обеспечение терминологической базы для идентификации патернов объектов предметной области (ПрО) предприятия в текстах и различных структурированных представлениях документов, например, таблицах, списках. В связи с чем, решение данной задачи в контексте интеграции различных корпоративных источников для организации эффективной работы информационного пространства связанных данных предприятия представляется нам актуальным и целесообразным.

На сегодняшний день информационное обеспечение предприятия часто строится на основе стихийной архитектуры, что подразумевает использование решений и приложений информационных систем от различных поставщиков. В такой архитектуре приложения интегрированы в негибкую инфраструктуру, в рамках которой не действуют унифицированные правила взаимодействия между системами, что приводит к необходимости создания интеграционных компонентов, которые должны быть достаточно адаптивными для

обеспечения эффективной и непрерывной работы предприятия [Завгородний, 2013].

В большинстве случаев предприятия имеют интеграционную архитектуру по топологии «точка-точка», внесение малейших изменений в которую может оказаться трудоемкой задачей, так как система сообщений и технологии, лежащие в основе, могут различаться.

Как правило, средние и большие предприятия состоят из территориально-удаленных подразделений, что требует использования эффективных средств коммуникации [Берко, 2009].

В последнее время, одним из перспективных направлений является построения эффективных систем электронного документооборота на основе слабосвязной архитектуры, что позволяет связать существующие задачи бизнес процессов предприятия единым программным обеспечением для минимизации затрат и увеличения степени интегрированности данных с возможностями организации открытого доступа к ним [Чистякова, 2014].

Целью данной работы является повышение эффективности систем электронного документооборота предприятий путем разработки модели и способа интеграции корпоративных данных на основе Wiki-систем и технологий связанных данных предприятия (англ. Linked Enterprise Data, LED) .

В данной работе в качестве источников данных будем рассматривать корпоративные документы, расположенные в Wiki-системах, а в качестве средства обеспечивающего терминологическую базу будем использовать корпоративные онтологии. Ограничение выбранного набора средств решения задачи ни в коей мере не сужает общности исследования данной работы в связи с их унифицированной природой.

1. Интеграционные технологии современного предприятия

1.1. Классификация технологий интеграции корпоративных данных

Основной целью предприятия является эффективная реализация бизнес-процессов для обеспечения максимальной прибыли. Для достижения этой цели разработан ряд способов и инструментальных средств описания, проектирования и анализа бизнес-процессов, в рамках которых технологии интеграции являются одними из важнейших составляющих. Таким образом, становится актуальным вопрос выбора технологии интеграции среди множества существующих. Рассмотрим виды, уровни и способы интеграции информационных систем, и определим наиболее эффективные интеграционные технологии для предприятий с территориально-распределенной структурой.

В настоящее время выделяют несколько видов интеграции информационных систем (рис. 1) [Франгулова, 2010].

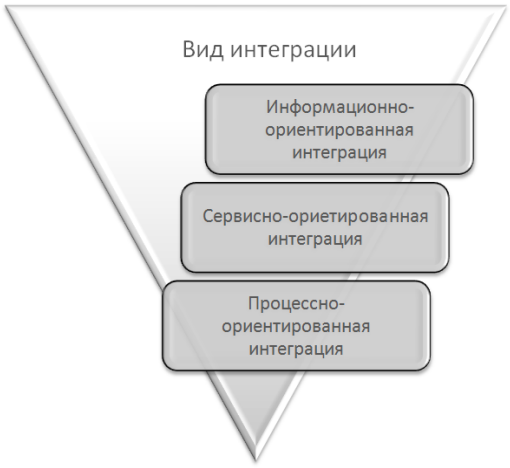

Рисунок 1 - Виды интеграции информационных систем

Информационно-ориентированная интеграция применяется, в основном, когда необходим обмен информацией между несколькими ИС. В процессе работы информационно-ориентированная интеграция использует обычно брокеры сообщений, связывающее программное обеспечение (ПО) (middleware), серверы репликации баз данных и другие технологии, целью которых является распространение информации между несколькими системами. Чаще всего данный вид интеграции используется при интеграции корпоративных приложений (Enterprise Application Integration, EAI) [Франгулова, 2010].

Технология сервисно-ориентированной интеграции опирается на слабосвязную архитектуру ИС, ориентированную на сервисы (Service Oriented Architecture, SOA). Эта технология применяется, когда необходимо совместное использование функций приложения и источников информации.

Сервисно-ориентированная архитектура, позволяет компоновать бизнес-процессы из компонентов, выполняющихся на разных платформах (корпоративных бинов J2EE, компонентов .NET, отдельных приложений), представлять в виде сервисов и повторно использовать в новых бизнес-процессах унаследованные компоненты [Гонтарь, 2013].

Процессно-ориентированная интеграция предоставляет возможность присоединиться к внутренним прикладным процессам каждого приложения, причем таким образом, чтобы не просто использовать его функции, а создать новый или мета-процесс, который и свяжет приложения [Франгулова, 2010]. Особенность использования этой технологии является возможность связывания большого числа разнородных информационных систем, используя при этом их встроенные функции.

В зависимости от подхода интеграции можно выделить следующие уровни: интеграция данных,

интеграция приложений, интеграция бизнес-процессов, интеграция на основе стандартов и интеграция платформ (рис. 2).

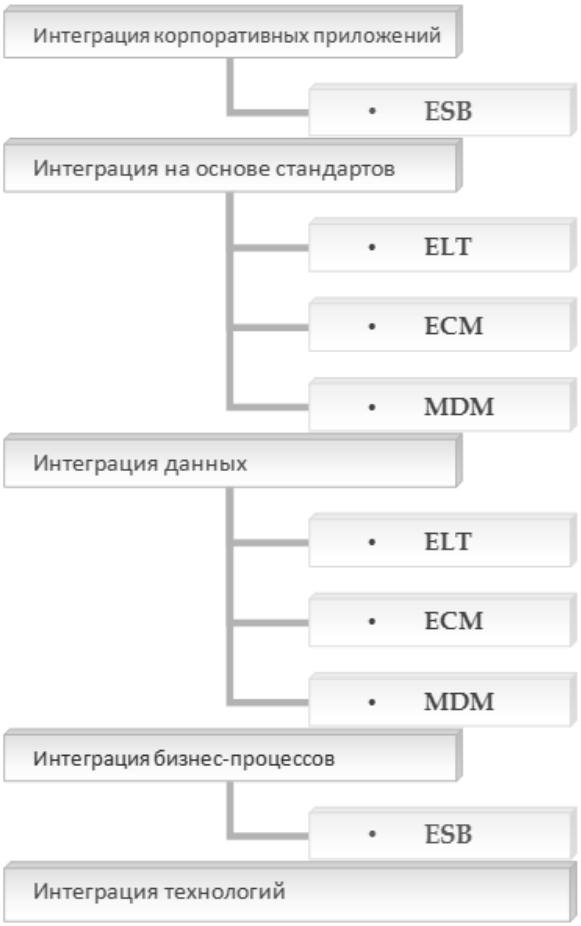

Рисунок 2 - Классификация интеграционных уровней и соответствующие им технологии

Интеграция бизнес-процессов (Business Process Integration, BPI) – основана на спецификации реализации и управления процессами обмена информацией между различными системами [Попов, 2013], что позволяет усовершенствовать операции интеграции и оптимизировать расходы в процессе использования ИС. Элементы BPI включают управление процессами, моделирование бизнес- и технологических процессов, которые охватывают различные задачи, процедуры, архитектуры, требования к входной и выходной информации, а также пошаговое разбиение каждого бизнес-процесса [Пушкарь, 2010].

Интеграция приложений (англ. Application Integration) осуществляется путем объединения данных или функций одной системы с другой. Передача функций или данных, свойственных какому-либо приложению, в распоряжение другого приложения используется с той целью, чтобы их взаимодействие обеспечило бы выполнение определенной прикладной функции ИС [Аткин, 2010].

Интеграция данных (англ. Data Integration) основана на идентификации и каталогизации данных с целью их дальнейшего использования. Успешная реализация интеграции бизнес-процессов и приложений на двух предыдущих уровнях зависит от того, как будут интегрированы в системе данные, принадлежащие разным источникам данных, в том числе баз и хранилищ данных. На этом уровне данные необходимо идентифицировать, каталогизировать и построить модель метаданных [Росинский, 2012].

Интеграция на основе стандартов (англ. Standards of Integration) основана на использовании стандартных форматов данных, таких как JSON, XML. В рамках такого подхода интеграция осуществляется на основе интеграционных схем на одном из языков описания данных, например, XML Schema [Арсеньев, 2001].

Интеграция платформ (англ. Platform Integration) касается процессов и инструментов, с помощью которых системы могут осуществлять безопасный и оптимальный обмен информацией. Обычно используется для завершения интеграции систем, формирования базовой архитектуры аппаратного и программного обеспечения и интеграции территориально-распределенных частей гетерогенной сети [Франгулова, 2010].

При межведомственной интеграции (англ. Interdepartmental Integration) заинтересованные в информационном обмене потребители заключают соглашения, определяющие состав данных, технологические, технические, организационные и экономические аспекты взаимодействия. В соответствии с регламентами ведомства передают часть своей информации в смежные ведомства, которые преобразовывают ее в соответствии с собственной терминологией и системой классификации. Часть информационных потоков являются однонаправленными, например, при формировании статистических отчетов), а часть – двунаправленными, ориентированных на взаимодействие запрос-отчет по запросу [Шаппелл, 2008].

Каждый подход имеет свои сильные и слабые стороны, но на практике используется комбинация различных подходов в зависимости от исходных условий и текущего уровня применения интеграционных средств и способов.

Проведём классификацию существующих способов интеграции.

По времени запуска:

1. Реального времени – если данные должны быть обновлены немедленно после изменений.
2. Отложенная – если процесс синхронизации данных запускается по какому-либо событию во времени или по расписанию.

По способу анализа информации:

1. По текущему состоянию – сравнение записей одной таблицы с записями другой, и на основании этого принимается решение о синхронизации,
2. Дельта-репликация – если в базе данных предусмотрен журнал вносимых изменений, и

алгоритм репликации переносит изменения по дельтам изменений, накопленным в журнале.

По направлению интеграции:

1. Односторонняя – если данные изменяются только в одном приложении, а в другой данные только хранятся и не подвергаются изменениям.
2. Многосторонняя – если данные могут изменяться и вводиться во всех приложения.

По уровню интеллектуального анализа:

1. Синтаксическая интеграция. Основывается на внешнем сходстве объединяемых данных.

Например, при объединении двух таблиц мы предполагаем, что в поле «№ договора» все записи имеют схожий формат «Договор № 3». Однако если в одной таблице в этом же поле дополнительно указывается дата договора «Договор № 1 от 17.01.2015», а в другой таблице выделено отдельное поле «Дата», необходимо обеспечить интерпретацию данных из одного вида в другой.

2. Семантическая интеграция. Основывается на сравнении данных на содержательном уровне.

Данный подход предполагает передачу вместе с данными также и их описание – метаданные. Этот тип интеграции основывается на знании и учёте природы данных. Основой семантической интеграции стала реализация онтологического подхода. При этом связь между элементарными единицами данных осуществляется в соответствии с их определением в корпоративной онтологии [Берко, 2009].

С учетом вышесказанного для интеграции корпоративных данных применим подход, ориентированный на использование корпоративной онтологии, в которой расположена большая часть доменной терминологии (терминологии ПрО), используемой предприятием.

1.2. Способ интеграции корпоративных данных в условиях слабой связности источников

Интеграция корпоративных данных заключается в последовательном выполнении нижеперечисленных этапов:

1. Формирование первичной системы типов связанных данных предприятия в виде корпоративной онтологии
2. Формирование первичной онтологической базы знаний предприятия в виде наборов квадов-четырехэлементных структур для описания графов связанных данных, путем составления синонимических рядов терминов, используемых на предприятии, и установления соответствия с типами корпоративной онтологии.
3. Подготовка типовых корпоративных документов и формирование на их основе терминологической базы путем выделения наиболее часто встречающихся терминов и выявления известных на основе корпоративной онтологии с учетом базы стоп-слов, исключающих добавление незначимых терминов.
4. Идентификация объектов в корпоративных документах и формирование схемы связанных данных корпоративного документа.
5. Формирование индекса связанных данных информационного пространства предприятия

Общая схема взаимодействия компонентов подсистемы интеграции системы электронного документооборота представлена на рис.3

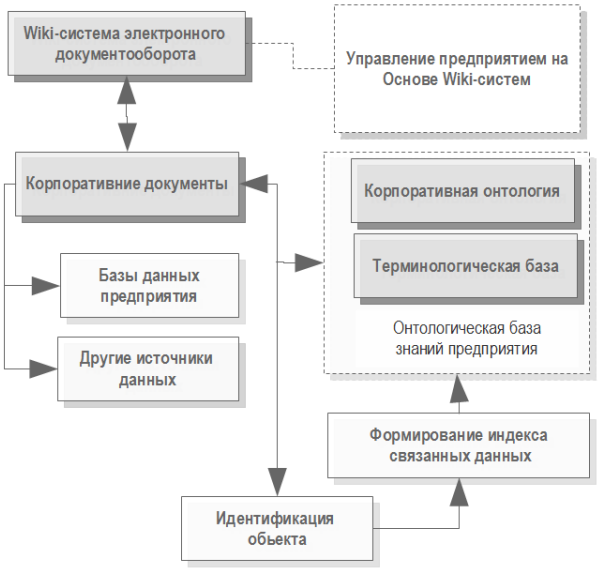

Рисунок 3 - Интеграция корпоративных данных

Рассмотрим более подробно составляющие этапы способа интеграции корпоративных данных.

На первом этапе администратор системы с специалистом предметной области формируют первичные знания о предметной области в виде системы типов связанных данных, определяя основные объекты и связи между ними.

На втором этапе формируется универсальное множество синонимов терминов, используемых в корпоративных документах, со спецификацией синонимичных рядов.

На третьем этапе, спецификации объектов и отношений схемы связанных данных используются для определения соответствующих им терминов и структурных образований, идентификация которых может быть осуществлена. В работе в качестве таких образований рассматриваются внедренные в документы таблицы и списки.

Далее, с целью определения положений этапа идентификации объектов, рассмотрим понятие корпоративного документа и разработаем соответствующую ему математическую модель.

Определение 1. Корпоративный документ – это иерархически организованный документ, представляющий собой множество вложенных заголовков, каждый из которых состоит из множества текстовых фрагментов, подзаголовков, таблиц, изображений, объединенных общим контекстом.

Замечание 1. Корпоративные документы относятся в большинстве случаев к классу слабоструктурированных документов.

Замечание 2. С учетом ориентации работы на использование результатов исследований в wiki-ориентированных системах электронного документооборота, то расширим набор используемых структурных компонентов семантическими аннотациями терминов, поддерживающимися расширениями Wiki-систем Semantic MediaWiki.

Формализуем понятие корпоративного документа.

Пусть Doc – корпоративный документ, H – множество вложенных заголовков, G – контекст документа, L – множество списков и таблиц документа, I – множество изображений, P – подзаголовки, W – семантические аннотации, то в соответствии с определением 1 с учетом замечания 1 и 2 корпоративный документ может быть представлен следующим выражением:

$$Doc = < H, G, L, I, P, W > \quad (1)$$

Замечание 3. В работе под понятием множества во всех случаях без исключения подразумевается конечное множество элементов.

Замечание 4. Корневой заголовок документа, даже при его отсутствии рассматривается в работе как существующий для организации единой точки доступа к содержимому документа при его обработке.

Замечание 5. Списки в работе рассматриваются как одноколоночные таблицы, что не сужает ни в коей мере общности рассуждений.

Таким образом, математическую модель корпоративного документа, согласно определения 1, с учетом замечаний 1-5 представим выражением 1.

На основе выше представленной модели корпоративного документа разработаем способ идентификации объектов в корпоративных документах.

Пусть $A = \{a_1, ..., a_n\}$ – терминологическая база предприятия, полученная в результате предобработки типовых документов предприятия и состоящая из n терминов, $B = \{b_1, ..., b_k\}$ – набор из *k*-стоп-слов, причем $A \cap B = \varnothing$, $S = \{S_1, ..., S_m\}$, где $S \subset S^U$ – синонимический ряд термина $a \in A$, S^U – множество всех синонимических рядов, тогда преобразование терминологических наборов в онтологическую базу знаний подразумевает установление соответствия между термином терминологического набора и типом онтологической базы знаний, что может быть выражено с помощью следующего отображения φ_1:

$$\varphi_1 : A \xrightarrow[S]{} T^U \quad (2)$$

Определим спецификацию типа объекта t, являющего элементом универсального множества типов T^U, используемого в выражении 2, с помощью следующего выражения:

$$\forall t \in T^U : t = < S, R, I > \quad (3)$$

где S – синонимический ряд термина, соответствующего типу объекта предметной области *t*, *R* – конечное множество отношений, характерных для данного типа, *I* – множество объектов типа *t*.

С учетом выражения 3 расширение синонимичного ряда соответствующего термину объекта представим как процесс добавления различных терминов с соответствующими им синонимическими рядами к объектам типа *t* с помощью выражения 4:

$$S^t = S^t \bigcup [S_1 \bigcup ... \bigcup S_p] \quad (4)$$

где S^t – синонимический ряд термина объекта t –го типа, причем $S^t \subset S \subset S^U$, S^U – универсальное множество синонимических рядов терминов предприятия.

В качестве программной реализации предложенных в статье модели и способов была использована система Semantic MediaWiki, а в качестве онтологической базы знаний предприятия использовался Openlink Virtuoso с встроенным RDF-хранилищем (англ. Resource Description Framework).

Заключение

В работе рассмотрены интеграционные технологии в контексте решения задач повышения рентабельности и эффективности электронного документооборота предприятия.

Рассмотрены технические аспекты совместного применения Wiki-систем с хранилищами триплетов RDF для организации электронного документооборота предприятий с территориально-распределенной структурой.

Получила дальнейшее развитие классификация подходов и способов интеграции корпоративных данных в рамках лоскутной автоматизации предприятий.

Предложена математическая модель корпоративного документа в виде графа связанных данных, которая в отличие от существующих поддерживает распределенное хранение и гипертекстовые семантические аннотации структурных компонентов, что позволяет обеспечить формальную основу для описания процесса интеграции корпоративных данных территориально-распределенных предприятий.

Получил дальнейшее развитие способ интеграции корпоративных данных в условиях структурной неопределенности и слабой связности источников, который в отличие от существующих позволяет добавлять новые источники данных без изменения общего алгоритма интеграции за счет

настройки критериев интерпретации структурных и семантических компонентов документов.

Предложен способ идентификации объектов в корпоративных документах на основе онтологического подхода, который позволяет формировать и расширять терминологическую базу в соответствии с системой типов, используемой на предприятии.

Рассмотрены практические аспекты реализации программных систем интеграции корпоративных данных на основе предложенных в работе моделей и способов.

Библиографический список

[Завгородний, 2013] Завгородний В. В. Информационная технология разработки специализированной СППР оперативного управления производством полупроводниковых изделий / В. В. Завгородний, И. В. Шевченко, В. Ф. Шостак, С. С. Щербак // Вісник Академії митної служби України. Серія: "Технічні науки". 2013 – № 1. – С. 69–76.

[Берко, 2009] Берко А.Ю. Способи та засоби семантичної інтеграції даних / А.Ю. Берко // Інформаційні системи та мережі. Вісник Національного університету "Львівська політехніка". 2009 – № 653. – С. 190–199.

[Чистякова, 2014] Чистякова И.С. Онтолого-ориентированная интеграция данных в семантическом вебе / И.С. Чистякова // Проблеми програмування. 2014. – № 2.–3. – С. 190–196.

[Франгулова, 2010] Франгулова Е. В. Классификация подходов к интероперабельности информационных систем / Е. В. Франгулова // Вестник АГТУ. Сер.: Управление, вычислительная техника и информатика. 2010. – № 2. – С. 176–180.

[Гонтарь, 2013] Гонтарь Н.А. Модель семантической сервис-ориентированной архитектуры / Н.А. Гонтарь // Наукові праці ДонНТУ. Сер.: "Інформатика, кібернетика та обчислювальна техніка". 2013 – № 1. (17) – С. 68–73.

[Попов, 2013] Попов В.А. Способ построения интегрированной системы управления предприятием на основе принципов непрерывного улучшения бизнес-процессов / В.А. Попов, А.В. Котляров // Авіаційно-космічна техніка і технологія. 2013– № 2. (37) – С. 144–151.

[Пушкарь, 2010] Пушкарь А.И. Изменения бизнес-процессов предприятий по созданию и реализации информационных продуктов и услуг в интернет-среде / А.И. Пушкарь, С.А. Назарова // Системи обробки інформації. 2010 – № 7 (88) – С. 167–173.

[Аткин, 2010] Аткин А. Интеграция ИТ: основные понятия и технологии / А. Аткин // Информационные технологии в экономике, управлени и образовании. 2010 – С. 284–289.

[Росинский, 2012] Росинский В.В. Обеспечение интеграции данных в корпоративных информационных системах на основе прогрессивных WEB-технологий / В.В. Росинский // Вісник ДУІКТ. 2012 – Т.10, №1. – С. 87–94.

[Арсеньев, 2001] Арсеньев Б. П., Яковлев С. А. Интеграция распределенных баз данных. – СПб.: Издательство «Лань», 2001. – 464 с.

[Шаппелл, 2008] Шаппелл Д. ESB – Сервисная шина предприятия: Пер. с англ. – Спб.: БХВ-Петербург, 2008. – 368 с.

[Лис, 2010] Лис К. П. Онтологическая интеграция данных моделирования для управления сервисно-ориентированной ИТ-инфраструктурой // Материалы 6-й международной конференции СпбГУЭФ. – Спб: Изд-во СпбГУЭФ. – 2010. – 62-67с.

[Андон, 2006] Андон П. Проблеми побудови сервіс-орієнтованих прикладних інформаційних систем в semantic web середовищі на основі агентного підходу / П. Андон, В. Дерецький // Проблеми програмування. 2006 – № 2.–3. – С. 493–502.

[Глибовец, 2013] Глибовец А.Н. Семантическая паутина и WIKI-системы / А.Н. Глибовец, Н.Н. Глибовец, Д.Е. Покопцев, М.О. Сидоренко // Проблеми програмування. 2013 – №1. – С. 45–67.

[Андон, 2006] Андон П. Проблеми побудо- ви сервіс-орієнтованих прикладних інформаційних систем в semantic Web середовищі на основі агентного підходу / П.Андон В. Дерецький // Проблеми програмування. – 2006. – № 2-3. – (спец. вип.). – С. 493–502.

[Боркус, 2006] Боркус Владислав. Способы и инструменты интеграции корпоративных приложений: Отчет/ RC Group.– М.: RC Group, 2006.– 13 с.

[Гудов, 2006] Гудов А.М. Интеграция распределённых приложений при помощи системы электронного документооборота. / А.М.Гудов, С.Ю. Завозкин // Труды международной конференции "Вычислительные и информационные технологии в науке, технике и образовании". II том – Павлодар: ТОО НПФ "ЭКО", 2006.

[Росинский, 2010] Росинский В. В. Способы и средства интеграции в CRM-системах / В. В. Росинский // Системные технологии. – 2010. – №6(71). – С.197–207.

[Вавилов, 2011] Вавилов К. П. Web-интеграция корпоративных систем / К. П. Вавилов. // Информационные технологии моделирования и управления. – 2011. – №3(68). – С.341-347.

ENTERPRISE DATA INTEGRATION METHODS UNDER CONDITIONS OF LOW SOURCES RELATEDNESS

Galushka I.M, Oksanich I.G., Shcherbak S.S

Kremenchuk Mykhailo Ostrohradskyi National University,

ilona.galushka@ya.ru

Technological aspects of the integrated use of Wiki-systems and RDF triplestore for organization of enterprise e-document workflow with geographically distributed structure were considered, existing approaches and methods for integrating enterprise data were analyzed and relevant classification was given.

Main Part

We propose a model of corporate document as the main source of linked data in business processes. We have developed a method for enterprise data integration based on the linked enterprise data concept and propose a method for identifying objects in corporate documents as a means of forming enterprise term base.

Integration technologies in the context of efficiency increasing of enterprise e-document workflow were considered.

Technological aspects of integrated use of Wiki-systems and RDF triple stores for organization of enterprise e-document workflow with geographically distributed structure were considered.

The classification of approaches and methods of enterprise data integration within patchwork automation of enterprises got further development.

Conclusion

The mathematical model of corporate document in the form of linked data graph was proposed; as opposed to the existing models, it supports distributed storage and hypertext semantic annotations of structural components that allows providing formal base to describe a process of enterprise data integration of enterprises with geographically distributed structure.

Practical aspects of enterprise data software systems implementation based on the proposed models and methods are considered.