Односторонняя интеграция информационных систем

Д. С. Порай, И. А. Тарханов

При создании информационных систем для крупных организаций часто встает задача интеграции новой системы с целым комплексом других систем, уже работающих на предприятии. Часто при этом затруднено или даже невозможно выполнить доработки этих систем с целью обеспечить возможность интеграции с новым программным комплексом. В данной работе введено понятие односторонней интеграции для решения подобных задач. Приведен пример решения такой задачи для интеграции Электронного Архива персонифицированного учета с Программно-техническим комплексом Система Персонифицированного учета для Пенсионного Фонда России.

1. Введение

В процессе увеличения масштаба, усложнения архитектуры, наращивания функциональных возможностей многие информационные системы (далее ИС) сталкиваются с проблемой интеграции с другими ИС [2]. Часто возникают ситуации, в которых несколько ИС становятся частью одного бизнес процесса, хотя изначально никакое взаимодействие между ними не планировалось.

В этой ситуации встает задача организации необходимого взаимодействия минимальными изменениями в существующих информационных системах. В данной статье приведена разработанная классификация ИС, классификация возможных типов взаимодействия между ними, описана реализация одного из типов взаимодействия на примере взаимодействия Электронного Архива и Системы Персонифицированного учета ПФР.

2. Классификация взаимодействующих информационных систем

При рассмотрении задачи интеграции разработанных в разное время информационных систем оказывается полезной классификация ИС по времени их создания и по степени поддержки:

1. Назовем ИС «перспективной», если возможности интеграции с другими ИС были заложены на этапе проектирования этой системы.
2. Назовем ИС «современной», если она уже внедрена в промышленную эксплуатацию, и ее разработка на момент постановки задачи завершена. Такого рода системы можно тоже разделить на два подтипа:
   2.1 ИС, в которых работы по изменению в принципе не проводятся. Как правило, это устаревшие программы, которые по каким-то причинам невозможно модифицировать (использование устаревшего языка программирования и/или СУБД, отсутствие разработчиков, изначально создавших ИС, отсутствие исходных кодов ИС и т. д.).
   2.2 ИС, для которых предполагается выпуск новых версий с расширенными функциональными возможностями.

3. Односторонняя интеграция информационных систем

Самая распространенная исходная ситуация – необходимо внедрить «перспективную» систему в уже существующий бизнес процесс предприятия. Ясно, что в «перспективную» систему изначально должны быть заложены возможности интеграции с существующим парком программного обеспечения с учетом того, на сколько возможно изменение уже существующих «современных» систем.

Если «современные» системы, участвующие в интеграции, не поддаются модификации или поддаются с трудом, то имеет место «односторонняя интеграция». Остановимся на этом понятии подробней.

Что означает термин «односторонняя интеграция»? Чаще всего это выражение встречается в новостях, когда она политическая сила предпринимает шаги к сближению, а другая нет. В случае информационных технологий картина похожая. Но в данном контексте вектор такой интеграции всегда или почти всегда идет от «современного» приложения к «перспективному». Каждая ИС имеет вход и выход в рамках общего бизнес процесса. Вход и выход неизменяемой «современной» системы тоже не меняется. Поэтому, в силу сложившихся обстоятельств, «перспективной» системе остается научиться работать с входом и выходом из системы «современной».

Во многих случаях этих входов и выходов оказывается недостаточно для реализации необходимых функциональных возможностей. Нужно внести изменения в «современную» ИС для того, чтобы у нее появились дополнительные входы и выходы. Здесь нужно четко понимать, какие изменения возможны в рамках рассматриваемой «современной системы»:

- Предоставить внешний интерфейс для доступа к функционалу «современной системы». Требуется изменение кода «современной» системы.
- Разработать дополнительный модуль для «современной» системы, который будет обмениваться данными с «перспективной» системой через специализированный формат обмена. Изменение кода самой «современной» системы, возможно, не потребуется, но необходима разработка дополнительного модуля.
- Реализовать в «перспективной» ИС прямой доступ к базе данных «современной» ИС. Изменение кода «современной» системы не требуется.

Каждый из способов имеет свои достоинства и недостатки. Первое и второе решение может оказаться невозможным в случае, если «современная» ИС в принципе не поддерживается. Третье решение имеет несколько аспектов:

- Оно может оказаться крайне сложным из-за того, что «современная» ИС использует устаревшую СУБД и/или устаревшую платформу (например, «современная» ИС использует СУБД типа Clarion или Clipper под MS-DOS, а «перспективная» разрабатывается под Windows с применением промышленной СУБД).
- Оно приводит к тому, что часть кода будет дублироваться – она будет реализована и в «современной», и в «перспективной» ИС. Эти две реализации могут оказаться не полностью идентичными, и тогда в базе данных «современной» ИС может возникнуть несогласованность.
- Оно требует решения вопроса с защитой данных в базе «современной» ИС. Может потребоваться создание дополнительных пользователей, создание представлений, расстановка прав доступа. Т. е. несмотря на то, что «современная» ИС в целом не модифицируется, на уровне базы данных может потребоваться внесение изменений.

Важно отметить, что третий подход может оказаться не только самым простым в реализации, но и самым эффективным, если необходимо только извлекать данные из базы данных «современной» ИС, и нет необходимости изменять эти данные (необходим режим «только для чтения»).

В каждом конкретном случае решение о выборе технологии интеграции должно приниматься после анализа всех упомянутых аспектов. Но во всех технологиях можно считать, что «современная» ИС приобретает дополнительные входы и выходы, которые необходимы для решения задач интеграции.

4. Пример односторонней интеграции

4.1. ЭАПУ и ПТК СПУ

Далее в качестве примера односторонней интеграции рассматривается встраивание «перспективной» системы «Электронный Архив документов индивидуального (персонифицированного) учета» (ЭАПУ) в бизнес процесс обработки документов персонифицированного учета Пенсионного Фонда России [1].

ЭАПУ предназначен для ввода, хранения, поиска, извлечения и печати документов персонифицированного учета, поступающих в бумажной и электронной форме.

До создания ЭАПУ процесс обработки документов персонифицированного учета выглядел следующим образом. Главным потребителем документов являлся программно-технический комплекс «Система персонифицированного учета» (ПТК СПУ). Документы поступали в одном из трех видов, каждый из видов обрабатывался отдельной цепочкой:

- Неформатные бумажные документы. В большинстве своем такие документы сопровождались файлом на дискете, который содержал в электронном виде данные, напечатанные на бумаге. Файл с дискеты загружался в ПТК СПУ, после чего производилась сверка бумажного оригинала с содержимым загруженного файла. В случае, если бумажный документ приходил без файла на дискете, содержимое документа набиралось оператором вручную.
- Формализованные бумажные документы на бланках. Документы проходили сканирование, распознавание и верификацию в «Автоматизированной системе массового ввода документов персонифицированного учета» (АСМВ СПУ). Результатом обработки являлся файл, который передавался в ПТК СПУ.
- Электронные документы в виде файлов с электронной цифровой подписью (ЭЦП). Файлы обрабатывались в Системе криптографической защиты информации (СКЗИ), в которой происходила проверка подлинности ЭЦП, и ее снятие с файла. Результатом обработки являлся файл, который передавался в ПТК СПУ.

Создание ЭАПУ внесло следующие коррективы в схему обработки документов (см. рис. 1):

- Неформатные бумажные документы стали передаваться в ЭАПУ. Там они сканируются, выполняется полуавтоматическое выделение значений реквизитов, изображения документов вместе со значениями реквизитов вводятся в базу данных.

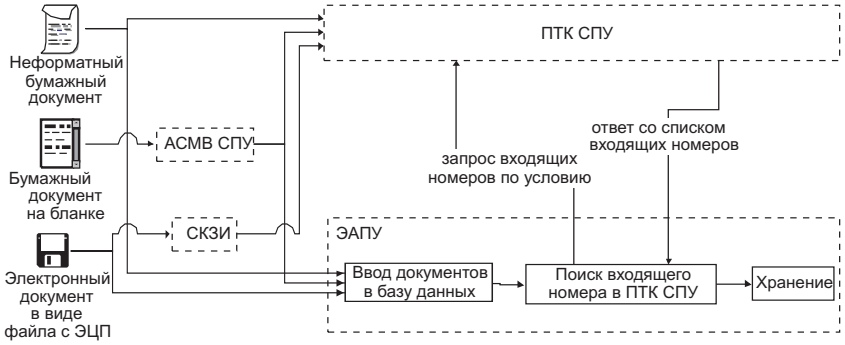

Рис. 1. Схема встраивания ЭАПУ в процесс обработки документов персонифицированного учета

- Формализованные бумажные документы на бланках передаются из АСМВ СПУ в ЭАПУ. Передаче подлежат изображения документов вместе с распознанными и проверенными значениями реквизитов. Вся эта информация вводится в базу данных.
- Электронные документы в виде файлов с ЭЦП передаются в ЭАПУ в оригинальном виде.
- В процессе обработки документов в ЭАПУ выполняется этап поиска входящего номера в ПТК СПУ. Этот этап является одним из важных пунктов требований к системе. Для однозначной идентификации документов ЭАПУ должен хранить уникальный входящий номер, присвоенный в ПТК СПУ при вводе документа.

В новой схеме обработки документов информационные системы классифицируются следующим образом:

- ЭАПУ — «перспективная» ИС.
- ПТК СПУ — «современная» ИС без возможности внесения изменений.
- АСМВ СПУ — «современная» ИС, новые версии которой регулярно выходят.
- СКЗИ — не связана с ЭАПУ, поэтому далее не рассматривается.

Из двух «современных» ИС АСМВ СПУ имела возможность экспорта изображений отсканированных документов и значений реквизитов. Для реализации интеграции потребовалась доработка, связанная с экспортом информации о многостраничных документах. Поскольку АСМВ СПУ является живой системой с регулярно выходящими новыми версиями, реализовать такую доработку не составило труда. От ЭАПУ

требовалось выполнить разбор файла в формате АСМВ СПУ и загрузить информацию в базу данных.

ПТК СПУ не включала в себя возможностей для обмена информацией о присвоенных входящих номерах. Новые версии ПТК СПУ выходят регулярно, однако эта система развивается независимо от ЭАПУ, и внести в нее дополнительные функциональные возможности оказалось невозможно. Поэтому с точки зрения ЭАПУ она является «современной» ИС без возможности внесения изменений. Для решения задачи интеграции было принято решение реализовать в «перспективной» ИС ЭАПУ прямой доступ к базе данных «современной» ИС ПТК СПУ. Анализ аспектов показал следующее:

- Обе системы реализованы на одной платформе — IBM eServer i5 (прежние названия IBM AS/400, IBM iSeries) с СУБД DB2/400. Обе системы физически работают на одном и том же сервере (региональный сервер в каждом субъекте РФ). Поэтому проблем с разнородностью платформ не возникло.
- ЭАПУ обращается к базе данных ПТК СПУ в режиме «только для чтения». Поэтому дублирующего кода не возникло, и несогласованность появиться в принципе не может.
- Вопрос с защитой данных в базе данных ПТК СПУ был решен следующим образом:
  - на сервере eServer i5 создан специализированный пользователь, этот пользователь находится в отключенном состоянии (DISABLED), что не позволяет войти в систему с его правами;
  - в базе данных ПТК СПУ созданы представления, такие, что выделенный пользователь имеет доступ только к ним из всей базы данных ПТК СПУ;
  - доступ к данным происходит исключительно через хранимую процедуру, запускаемую с правами выделенного пользователя, ни один пользователь ЭАПУ не имеет прямого доступа к созданным представлениям.

Рассмотрим далее более подробно решение задачи интеграции ЭАПУ и ПТК СПУ для операции «Поиск входящего номера» с учетом выбранного способа интеграции двух систем. Следует отметить, что это не единственная операция, в которой есть взаимодействие двух систем — есть еще синхронизация нескольких справочников. Однако эти операции являются тривиальными и здесь не рассматриваются.

4.2. Представление данных в ПТК СПУ

Интеграция ЭАПУ и ПТК СПУ основана на следующих основополагающих принципах:

- Существует более десяти типов документов персонифицированного учета.
- Все документы обрабатываются пачками.
- Первым документом в каждой пачке является опись (обычно, но не всегда). Опись хранит основную информацию о пачке, включая тип документов в пачке и количество документов.
- Все остальные документы в пачке имеют один и тот же тип.
- Описи соответствует представление. Каждому типу документов также соответствует представление.
- Ключевым полем в базе данных ПТК СПУ для всех типов документов (включая опись) является входящий номер, который состоит из кода подразделения (района), входящего номера пачки по данному подразделению, номера документа в пачке.
- В базе данных ЭАПУ код подразделения, входящий номер и номер документа не являются ключевыми полями при организации базы данных, но их уникальность проверяется при вводе документов в Архив.

4.3. Проблемы и требования организации поиска

Несколько особенностей хранения пачек в ПТК СПУ существенно усложняют задачу интеграции:

1. Для некоторых типов документов (например, ведомость уплаты страховых взносов) информация об описи отсутствует в базе ПТК СПУ.
2. Возможны случаи, когда в документном представлении отсутствует информация о пачке совсем или об отдельных документах пачки.
3. Нет точного соответствия представления в базе данных и типа документа в пачке. Например, информация о документах из пачки индивидуальных сведений может содержаться как в одном из двух представлений («старые» и «новые» индивидуальные сведения).
4. Существует трудность в установлении уникальности записи по документным представлениям. Каждое из представлений имеет свой набор полей и особенности их заполнения.
5. Существует семь версий форматов файлов персонифицированного учета (четыре текстовых формата и три XML формата). При обращении к представлениям ПТК СПУ необходимо разбирать все типы документов во всех форматах.
6. Проблема производительности. В региональных базах ПТК СПУ представления могут содержать по несколько десятков тысяч записей, по этому «цена» каждого запроса к базе ПТК СПУ очень высока.

7. Неоднородность региональных баз ПТК СПУ. На некоторых региональных серверах отсутствуют представления для некоторых типов документов.

Из описания проблематики и общего устройства представлений ПТК СПУ видна сложность однозначного поиска входящего номера для пачки при вводе в Архив.

Отсюда, вытекает несколько очевидных требований по организации поиска:

1. Согласно пунктам 1 и 2 предыдущего списка, нужно осуществлять отдельно поиск по таблице описей и отдельно по таблице документов.
2. Согласно пунктам 3 и 4 предыдущего списка, необходимо параметризовывать имена, количество представлений и поля в них для поиска документов из пачки.
3. Пункт 5 предыдущего списка накладывает необходимость более детального разбора и доступа к информации из файлов с ЭЦП всех форматов.
4. Последние два пункта заставляют обязательно проверять, существует ли необходимое представление и делать запрос к нему максимально оптимизированным.

4.4. Параметризация поиска входящего номера

Необходимость гибких настроек при поиске и его параметризация была очевидна уже при постановке задачи односторонней интеграции ЭАПУ и ПТК СПУ. В ЭАПУ для этого используется текстовый INI-файл.

Во-первых, в представлении описи для каждого типа документов есть соответствующее поле, значение которого — количество документов в пачке. Для учета этого при поиске описей пачек в файле заведена секция, в которой данные хранятся в следующем виде:

$$\{тип\_документа\} = \{имя\_поля\_в\_представлении\_описей\}$$

На основе этого параметра формируется запросы для поиска по описям, все ниже описанные параметры служат для формирования поиска по документам.

Искать все документы в пачке (их количество может достигать 200 штук) накладно. Поэтому, при большом количестве документов в пачке используется только выборка из нескольких штук. Размер выборки для каждого типа документов в пачке параметризован:

$$\{тип\_документа\} = \{размер\_выборки\_для\_поиска\_документов\}$$

Хранится информация, в каких документных представлениях ПТК СПУ нужно искать документы определенного типа. В соответствующей секции файла это сделано в виде записей:

$$\{тип\_документа\} = \{имя\_ представления1\_для\_поиска\_документов\}/ \{имя\_ представления2\_для\_поиска\_документов\}/...$$

Кроме названия представления для генерации текста запроса необходимо знать название полей в этом представлении. По условию из комбинаций значений этих полей должен однозначно находиться документ в пачке. Назовем такое поле «уникальным полем», а его значение «уникальным значением» этого поля.

$$\{имя\_представления\_для\_поиска\_документов\} = \{имя\_уникального\_поля1\}/\{имя\_уникального\_поля2\}/...$$

В качестве уникальных полей при поиске используются регистрационные номера страхователей, отчетные периоды, различные даты. Реже имена, фамилии, отчества, страховой номер, суммы.

Так же в качестве параметра в отдельной секции указывается тип поля и его нахождение в файле с ЭЦП определенного формата. Этот параметр используется для дополнительного разбора файла с ЭЦП.

4.5. Алгоритм поиска в базе ПТК СПУ

Рассмотрим алгоритм поиска входящих номеров в доступных представлениях базы данных ПТК СПУ. Логика, описываемая ниже, реализована частично в SQL-запросах, частично в C++ коде клиентской части при анализе результатов этих запросов.

Мы имеем два типа специализированного поиска, дающие одинаковый результат — уникальную пару (dpt — код района, dci — входящий номер), — поиск по описи ($O(dpt, dci)$) и поиск по документам пачки ($D(dpt, dci)$). Каждый результат из любого из этих множеств обладает вероятностью. Вероятность поиска по описи:

$$P_O = \sum_{j=1}^{j=n} K_j \cdot P_j,$$

где K_j — коэффициент j-го уникального поля из описи; $P_j = 1$, если значение поля j результата совпало со значением из файла с ЭЦП, и равно 0 — если нет.

Например, таким уникальным полем для описи является регистрационный номер. Коэффициенты подбирались опытным путем.

Видно, что вероятность равна 1, если все поля в описи документа найдены в полях результата.

Таблица 1

Уникальные поля для поиска входящих номеров по описи

Поле из описи	Коэффициент
Регистрационный номер страхователя	0,3
Тип документа пачки	0,5
Номер пачки	0,1
Количество документов в пачке	0,1

Таблица 2

Значение коэффициента размера выборок для каждого типа документов

Тип документа	Значение коэффициента
Индивидуальные сведенья (до 200 документов в пачке)	7
Анкеты сведенья (до 200 документов в пачке)	7
Ведомость уплаты страховых взносов (обычно 1 документ в пачке)	1
Сводная ведомость (обычно 1 документ в пачке)	1
Заявление об обмене страхового свидетельства (обычно 5–10 документов в пачке)	3

Вероятность поиска по документам должна, в свою очередь, учитывать информацию обо всех документах, входящих в данную пачку. Поэтому имеет смысл подсчитать количество точных совпадений по уникальным полям из представлений документов. Но здесь мы сталкиваемся с другим ограничением, накладываемым количеством документов в пачке — оно может достигать 200. Поэтому используется коэффициент для каждого конкретного типа документов, показывающий количество документов в пачке для поиска. Этот коэффициент (K), упомянутый выше в разделе параметризации, ограничен сверху количеством документов в пачке. Значение его линейно зависит от среднестатистического количества документов в пачке.

Вероятность для результата поиска по документным представлениям:

$$P_d = \frac{\sum_{i=1}^{i=K} \left(\prod_{j=1}^{j=n} P_{ij} \right)}{K},$$

где n — количество уникальных полей документа; $P_{ij} = 1$, если значение уникального поля j результата i совпало со значением из файла с ЭЦП, 0 — если нет.

Таким образом, в простейшем случае, когда информация об описи пачки и документах присутствует всегда, окончательный результат это пересечения двух множеств:

$$R(dpt, dci) = O(dpt, dci) \cap D(dpt, dci),$$

с вероятностью P_d — как более верной с математической точки зрения.

В остальных случаях, если $O(dpt, dci) = \varnothing$, то

$$R(dpt, dci) = D(dpt, dci);$$

если $D(dpt, dci) = \varnothing$, то

$$R(dpt, dci) = O(dpt, dci).$$

Остается один теоретически возможный вариант, когда $O(dpt, dci) \cap D(dpt, dci) = \varnothing$, при $O(dpt, dci) \neq \varnothing$ и $D(dpt, dci) \neq \varnothing$, тогда

$$R(dpt, dci) = O(dpt, dci) \cup D(dpt, dci)$$

— это множество допустимых вариантов, которое должно быть проконтролировано пользователем ЭАПУ, как внештатная ситуация.

5. Заключение

В данной работе создана классификация взаимодействующих информационных систем. Введено понятие односторонней интеграции. Сформулированы возможные технические решения задачи односторонней интеграции. Перечислены достоинства и недостатки каждого решения.

Рассмотрен пример решения задачи односторонней интеграции для систем ЭАПУ и ПТК СПУ. Приведен разработанный в процессе решения данной задачи алгоритм сопоставления данных из двух БД. Данный алгоритм был разработан с учетом специфики хранения данных в базе ПТК СПУ и используется для поиска входящих номеров и кодов района для любых пачек, как автоматически, без вмешательства пользователя системы ЭАПУ, так и вручную, когда весь список найденных результатов показывается пользователю. Алгоритм также работает с более мягкими

условиями на проверку полей описи (например, не учитывается номер пачки). Логика работы при этом остается прежней.

Работа данного инструмента была проверена на промышленной базе ПТК СПУ во время опытной эксплуатации системы, в процессе которой подбирались коэффициенты и поля, множество которых уникально для каждого документа, для всех типов документов пачек персонифицированного учета.

ПТК СПУ является для ЭАПУ информационной системой, предоставляющей свои справочники страхователей, подразделений, отчетных периодов, а так же всю информацию о входящих номерах и кодах подразделений пачек. С помощью односторонней интеграции удалось осуществить косвенный контроль вводимых в Архив данных. При разработке ЭАПУ осталось лишь грамотно распорядиться представленными данными для решения поставленных при разработке системы задач, что и было сделано. Отметим, что поиск входящего номера в ПТК СПУ является основной операцией перед вводом в Архив, без которой дальнейшая работа ЭАПУ теряет всякий смысл.

Литература

1. *Порай Д. С.*, *Порай Т. А.*, *Соловьев А. В.* Построение расширяемого программного комплекса // Сборник трудов ИСА РАН. М.: УРСС, 2005.
2. *Тарханов И. А.* Интеграция Системы Электронных Выплатных Дел со сторонними приложениями // Сборник трудов ИСА РАН «Системный подход к управлению информацией». М.: УРСС, 2006. Т. 23.
3. *Порай Д. С.*, *Соловьев А. В.*, *Корольков Г. В.* Реализация концепции темпоральной базы данных средствами реляционной СУБД // Сборник трудов ИСА РАН «Документооборот. Концепции и инструментарий», 2004.
4. Постановление Правления Пенсионного Фонда Российской Федерации от 31.07.2006 № 192п «О формах документов индивидуального (персонифицированного) учета в системе обязательного пенсионного страхования и инструкции по их заполнению.» http://www.pfrmsk.ru/pu/doc/192p.doc
5. Материалы сайта http://www.opfr34.ru/sitepfr_1.5.1.htm
6. Аналитический обзор материалов журнала Data Communications International http://www.citforum.ru/nets/digest/dig_1903.shtml#10