А. Ю. Медведев

Концептуальные основы электронного сбора первичных данных

Рассматриваются концептуальные основы построения технологии ведения и передачи первичных данных с позиции интеллектуальных информационных систем.

Введение

Электронный документооборот вновь стал одной из популярных и актуальных тем научно-технических изданий по информационным технологиям в России [1–4]. В журнале «Открытые системы» анализируются многие его аспекты: жизненный цикл документа в системе электронного документооборота (СЭД) [1], виды хранилищ документов [1], выгодность внедрения СЭД [2], необходимость использования средств сетевой защиты электронного архива и разграничения прав доступа к нему [3] и др., однако не рассматривается важнейший этап электронного документооборота, связанный со сбором большого массива первичных данных. На наш взгляд, именно данный этап образует функциональную основу замкнутой системы государственного регулирования при контроле за работой большого количества подотчетных субъектов. Например, для эффективного государственного управления в налогообложении, образовании, здравоохранении, экологии и других сферах документооборот должен обладать такими качествами, как полнота, достоверность и оперативность информации, что может быть обеспечено полномасштабным применением современных информационных технологий. Поэтому в последнее время наблюдаются тенденции информатизации государственного управления [5].

Несмотря на существенные сдвиги в вышеуказанной области, реализация задачи сбора первичных данных в электронном виде органами государственного контроля от подотчетных субъектов выявляет определенные проблемы, в частности, при вовлечении огромного количества людей в сферу электронной информатики возникает проблема корректного сопровождения первичных данных. Под первичными данными будем понимать наиболее детальную информацию о контролируемом процессе, вносимую конечным пользователем*)*) в систему.

В статье рассматриваются концептуальные основы технологии ведения и передачи первичных данных в условиях действия следующих осложняющих факторов:

- необходимость приема, обработки и анализа больших массивов исходной информации;
- неопределенность и неполнота входных потоков данных;
- потребность в технологиях распознавания исходных образов и данных.

В научно-технической литературе [2, 6] информационные технологии, работающие в условиях действия осложняющих факторов, именуются интеллектуальными.

Важно подчеркнуть, что целью электронного сбора являются создание и обновление структурированной базы данных, способной к любым многовариантным анализам, именно это позволит принимать оптимальные решения на любом иерархическом уровне управления (например, городском, районном, областном, федеральном). Поэтому перевод первичных данных в электронную форму обязательно должен осуществляться по определенной модели, тре-бующей «жесткой» формализации документов. Нельзя смешивать его с обычным внесением информации в текстовых редакторах. Никакой архив Word- или Excel-документов не способен обеспечить анализа, это лишь дублирование бумажных форм, без возможности автоматического сбора данных.

В связи с ведением данных по единому электронному стандарту требования к внесению информации становятся более строгими. Электронный способ обязывает конечных пользователей использовать «зашитые» в программу справочники (при необходимости их оперативного обновления), невнесение нерегламентированных данных и пояснений не допускается (в отличие от бумажной формы), кроме того, на формализованном шаблоне отсутствует привычная для бумаги свобода внесения информации. Переход от ручных согласований принимаемых первичных данных в бумажных формах к автоматизированным может и должен привести к детализации собираемой информации, так как это повысит потенциальные возможности анализа и не приведет к усложнению приема. Электронная технология позволяет собирать уже не только ключевые сводные показатели, но и исходные данные, следовательно, подразумевает изменение стандартов передаваемых первичных форм. Это увеличивает трудоемкость внесения первичной информации и ужесточает регламент документооборота (требуется полное и своевременное внесение детализированных данных).

При малом числе конечных пользователей (например, использование тех-нологии внутри небольших предприятий) перечисленные выше особенности сбора первичной информации не имеют существенного значения. Государственное использование технологии, например в области налогообложения или экологии, подразумевает охват числа конечных пользователей, соизмеримого с количеством зарегистрированных в стране предприятий. В этом случае любое усложнение, изменение регламента способно вызвать долговременную адаптацию к нововведениям. Важно добиться компромисса между разумной детализацией собираемой первичной информации и увеличением трудоемкости ее ведения. В связи с этим разумно предложить технологию сбора первичных данных в электронном виде с надстройками интеллектуальных

информационных систем и применением гибких интерфейсов [6].

Организационная схема

Первоначальным этапом сбора данных в электронной форме является перевод бумажных отчетных форм в электронный формат. Этот этап выполняют подотчетные субъекты, внося первичные данные с помощью специализированных средств ввода. Следующий этап — передача электронных версий отчетности органу государственного контроля, она осуществляется в так называемый информационный буфер с помощью какого-либо канала связи (электронная почта, Интернет, гибкие магнитные носители). Далее орган контроля выполняет процедуру согласования данных (информация диагностируется на ошибки заполнения, достоверность и полноту), после чего согласованные данные вносятся в виртуальное хранилище [7], а несогласованные остаются в информационном буфере, формируя запрос на исправление данных (рис.).

Остановимся на ключевых моментах технологии подробнее.

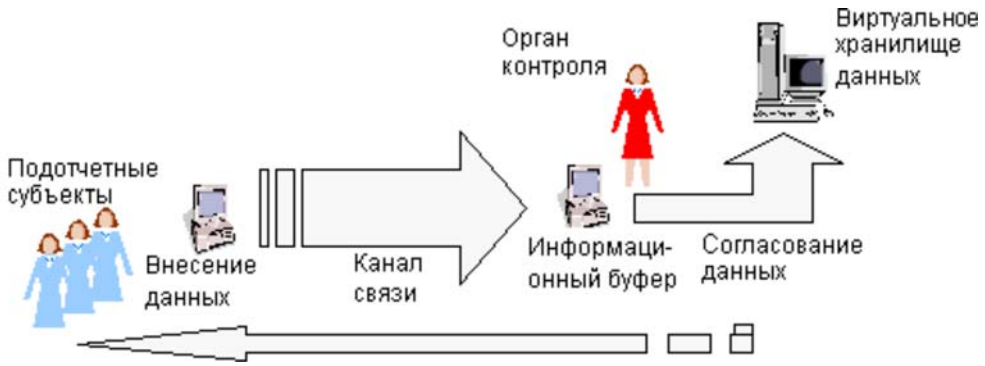

Рис. Организационная схема технологии электронного сбора информации

1. Единый электронный стандарт передаваемой информации. Информация, передаваемая в электронном виде, должна быть предварительно формализована в соответствии с разработанным стандартом. Формализация данных (структурирование, задание жестко определенных разделов, граф, строк с соответствующими взаимосвязями) крайне важна, так как электронный сбор информации в неформализованном виде (например, в произвольно заполненных документах MS Word или MS Excel) мало чем отличается от обычной бумажной отчетности — оперативность анализа и принятия управленческих решений в этом случае недостижима.

Важнейшей составляющей успеха сбора информации в электронном виде является грамотно разработанный единый стандарт передаваемой информации. Следует отличать электронный стандарт передаваемой информации от образцов передаваемых отчетных документов в бумажном виде (отчетных форм в определенном визуальном стандарте). Под электронным стандартом понимается специально разработанная структура базы данных принимаемой информации (описание структуры таблиц передаваемых данных с типизацией их полей, взаимосвязи таблиц). При разработке стандарта важно учитывать не только грамотность построения структуры БД с технической точки зрения, но и применимость его для отображения информации в привычном для отчитывающихся лиц визуальном стандарте. Желание отказаться от привычного визуального (ненормализованного) стандарта и визуализировать экономичный электронный стандарт (например, убрать сводные и оставить только исходные, т. е. первичные, данные) приведет к неузнаваемости отчетной формы отчитывающимся лицом, что затруднит заполнение отчета и может в целом затормозить внедрение электронной технологии.

2. Средство внесения электронных данных. Для перевода первичной информации в электронный вид важно разработать средство внесения информации, которое бы не только отображало вносимые данные в привычном образе отчетного документа, но и диагностировало их на предмет ошибок. Оно должно быть передано (или сообщен адрес его ресурса в Интернете) каждому подотчетному субъекту. В качестве такого средства может выступать разработанный шаблон внесения информации или специально разработанный программный модуль.

Шаблоны могут быть различных типов: определенно структурированный документ MS Word или MS Excel, XML-документ. В большинстве случаев предоставления статистической информации наиболее предпочтительны так называемые «легкие» шаблоны, т. е. не требующие специально разработанных программных средств и обучения отчитывающихся лиц дополнительным программам (которые к тому же обычно нуждаются в корректной установке и сопровождении). При работе с «легкими» шаблонами информация вносится в уже хорошо изученные программы, являющиеся неотъемлемой частью любого компьютера.

Однако использование «легких» шаблонов не всегда удобно. В случаях, когда необходимо передавать большой массив данных, лишь немного изменившихся по сравнению с прошлым отчетным периодом, более

приемлемо для подотчетных субъектов специализированное программное средство, в ко-тором передаваемая информация в межотчетный период может нужным образом сопровождаться, а в момент отчета просто предоставляться. Например, для создания так называемой статистической формы РИК-83 [4] в районных образовательных органах с каждого образовательного учреждения ежегодно собирается реестр преподавателей с информацией о профессиональной деятельности каждого из них. Учитывая, что это довольно большой массив данных, нецелесообразно ежегодно практически полностью дублировать его «вручную». Гораздо удобнее в подобных случаях иметь специализированные програм-
мные средства, которые могут автоматически формировать заполненный «легкий» шаблон (например, в XML-документе) из непосредственно прошлогодних или предварительно измененных прошлогодних данных для последующей передачи его в контролирующий орган.

Так как компьютеры отчитывающихся организаций могут управляться раз-личными операционными системами, предпочтительнее в качестве шаблона использовать XML-документ, который можно просмотреть и заполнить в любом html-браузере. Многоплатформенность XML-документа и возможность передачи его через Интернет непосредственно в базу данных контролирующего органа делает этот способ наиболее выгодным.

3. Канал связи. Заполнив шаблон, подотчетный субъект должен передать его контролирующему органу (принимающей стороне). Шаблон можно принес-ти на легких магнитных носителях (дискета, флеш-карта), переслать по электронной почте или Интернету (если шаблон представлен XML-документом). Кроме того, контролирующий орган также должен иметь возможность связи с от-
читывающейся стороной для передачи ей обновленных справочных и вспомогательных архивов.

Рассмотрим достоинства и недостатки каждого из способов.

Гибкие магнитные носители. Достоинство этого способа передачи электронных данных — его доступность. В отличие от электронной почты или Интернета, он не требует специального сетевого и коммутационного оборудования. Недостатки этого способа также очевидны — необходимо личное сопровождение дискеты до принимающей стороны при каждой посылке (как в момент первого согласования отчета, так и при последующих согласованиях, если выявлены ошибки).

Электронная почта. Достоинство этого способа по сравнению с предыдущим — возможность посылки сформированных электронных данных без личного участия отчитывающегося лица. Недостаток в том, что данные поступают не непосредственно в информационный буфер приемной стороны, из которого осуществляется процедура согласования, а в промежуточное информационное пространство, из которого орган контроля должен извлечь их для последующей пересылки в информационный буфер, чтобы согласовать и передать данные в общее виртуальное хранилище.

Внесение отчетных данных On-Line (через Интернет). Использование так называемого тонкого клиента [8] для интерактивного представления отчета непосредственно в информационный буфер контролирующего органа делает этот способ наиболее комфортным. В данном случае подотчетному субъекту нет необходимости беспокоиться о своевременности получения его данных приемной стороной, как при передаче дискеты (может оказаться негодной) или посылке по электронной почте (может долгое время не проверяться). Данные будут непосредственно внесены в информационный буфер, и останется лишь дождаться результата их согласования. При этом способе результаты согласования удобно разместить также в Интернете на сайте контролирующего органа. Недостатки указанного способа в том, что, во-первых, не у всех подотчетных субъектов есть свободный выход в Интернет, во-вторых, требуются значительные затраты со стороны контролирующего органа на внедрение на приемной стороне технических средств по комплектации и конфигурированию сетевого оборудования, средств защиты информации от несанкционированного взлома, содержание соответствующих специалистов.

4. Система приема электронных данных. На приемной стороне вся предоставляемая отчетная информация проходит стадию согласования и передачи в общее виртуальное хранилище данных. Для этой цели необходим специально разработанный программный модуль приема и согласования принимаемых данных. Отчетная информация должна диагностироваться на полноту внесения, «подозрительные» и ошибочные данные — автоматически выявляться. Важно, чтобы система позволяла накапливать несогласованные данные в информационном буфере, из которого после согласования они поступят в общее хранилище. Ответственность за передаваемую информацию лежит на отчитывающихся лицах, поэтому вся принимаемая информация должна содержать сведения об ее авторе.

Методы обеспечения полноты и целостности входного потока данных

Рассмотрим факторы, обусловливающие отсутствие полноты и целостности собираемых первичных данных, и предложим способы их устранения.

· Неориентированность на работу в информационных технологиях большого количества подотчетных субъектов и как следствие неверное (ошибочное) внесение первичной информации.

Устранение проблемы: создание простых интуитивно-понятных инструментов внесения первичной информации, включающих в себя анализатор вводимых данных на ошибки.

· Низкая достоверность собираемых данных из-за проблематичности оперативного обновления передающей стороной первичных данных, а принимающей — справочных и вспомогательных архивов.

Устранение проблемы: применение быстрых каналов связи для приема-передачи первичной информации

(электронная почта, Интернет); технологии тонкого клиента, позволяющей отчитывающейся стороне передавать данные непосредственно в БД принимающей стороны, пользуясь при этом актуальными справочными и вспомогательными архивами.

· Недостаточность полноты принятой первичной информации из-за отсутствия ряда отчетов или частичного заполнения отчетов (выборочного заполнения разделов отчета).

Устранение проблемы: возможность эмулирования непринятой первичной информации (создание недостающих отчетов и довнесение в незаполненные разделы правдоподобных данных) для более достоверного анализа данных. Эмулирование можно производить автоматической генерацией данных из прошлогодних или других правдоподобных источников.

· Прием первичных данных с ошибками уменьшает достоверность информа-ции и может вызвать критические ситуации при последующем анализе данных.

Прежде чем предложить методы устранения этой проблемы, классифицируем ошибки принимаемой информации.

Выделим два класса ошибок: *опознаваемые* и *не опознаваемые*.

Под опознаваемыми будем понимать явные ошибки внесения данных, которые можно автоматически распознать программными средствами; под не опознаваемыми — неверные, но правдоподобные данные, которые не подлежат выявлению программными средствами в автоматическом режиме.

Опознаваемые ошибки разделим в свою очередь на две категории: *односложные* и *многосложные*.

Под односложной будем понимать ошибку одного конкретного значения, не связанного с другими значениями передаваемого отчета; под многосложной — ошибку, связанную с нарушением определенно заданных взаимосвязей данного значения отчета с другими значениями первичной или справочной ин-формации.

Типы односложных ошибок:

· выход за пределы диапазона значений;
· неправильность типа введенных данных;
· лингвистическая ошибка;
· неверная точность вещественного значения;
· отсутствие значения;
· нарушение целостности данных;
· очевидное смысловое искажение.

Типы многосложных ошибок:

· несоблюдение баланса данных в строке (столбце) таблицы (часто различные значения в таблице взаимосвязаны по определенному принципу: балансовое соотношение значений строки или какие-либо другие контрольные суммы);

· несоблюдение баланса данных различных разделов шаблона (данный тип ошибки рассматривается отдельно в связи с тем, что иногда взаимосвязи разделов гораздо меньше диагностируются на предмет их нарушения по сравнению с диагностикой ошибок внутри одного конкретного раздела);

· нарушение определенно заданной взаимосвязи значения текущего отчета со значением другого отчета (подобные ошибки возникают, например, когда в текущий отчет внесены данные за прошлый год, не соответствующие данным прошлогодней отчетности);

· ошибки, связанные с использованием устаревших справочных архивов или их некорректными аналогами (возникают, когда обновление справочников на приемной стороне не синхронизировано с таковым на передающей стороне или когда последней предоставлена свобода корректировки либо довнесения справочных архивов).

Наряду с рассмотренными классами ошибок (опознаваемые и не опознаваемые), выделим еще два класса, характеризующих возможный способ исправления ошибок: *автоисправляемые* и *требующие ручного способа исправления*.

Под автоисправляемыми будем понимать такие опознаваемые ошибки, которые могут быть исправлены в автоматическом режиме на стадии ввода или диагностики информации. Например, неверный символ сепаратора в десятичной дроби или неверная точность вещественного значения (неверное число знаков после запятой) и др.

Под требующими ручного способа исправления будем понимать все ошибки за исключением автоисправляемых.

Используя приведенную терминологию, предложим некоторые методы из-бавления от ошибок в принимаемой первичной информации.

1. Проверка на ошибочность вводимых данных непосредственно при вводе и исправление автоисправляемых ошибок; индикация ошибок, требующих ручного способа исправления; автоматизация внесения суммарных значений отчета.

2. Анализ только что заполненного первичного отчета специальной диагностической программой-анализатором, выявляющей опознаваемые ошибки заполнения как на передающей, так и на принимающей стороне.

3. Применение тонкого клиента как средства внесения информации, что позволит исключить использование неактуальных или некорректных справочных архивов. При использовании других средств

внесения первичной информации необходимы создание инструмента корректного обновления справочников и согласование всех «неизвестных» введенных в справочный и вспомогательный архивы значений.

4. При появлении на стадии анализа сомнительных или явно ложных данных (после исключения опознаваемых ошибок на стадии согласования данных) необходим анализ первичной информации на выявление не опознаваемых ошибок. Для этого следует существенно сузить список отчетов, подлежащих досмотру, так как «ручная» проверка всех отчетов в силу их огромного количества просто невозможна. Для сужения списка должны быть разработаны критерии выбраковки наиболее сомнительных отчетов, основанные на предметной области проводимой отчетности.

Литература

1. *Пахчанян А.* Технологии электронного документооборота // Открытые системы. 2002. № 10. С. 17–21.

2. *Головко М.* Технический документооборот: Система управления документацией или придаток к приложениям обработки данных? // Там же. С. 22–25.

3. *Климов В.*, *Краюшкин В.*, *Пирогова М.* Архив на базе PDM // Там же. С. 26–33.

4. *Бакланов А. В.*, *Долгушин В. В.*, *Сердитов Ю. Г.*, *Соловьев И. Г.* Технология сбора образовательной статистики на основе электронных шаблонов // Вестн. кибер-нетики. 2004. № 3. С. 42–49.

5. *Концепция* использования информационных технологий в деятельности федеральных органов государственной власти до 2010 года. Одобрена распоряжением Пра-вительства Российской Федерации от 27 сент. 2004 г. № 1244-р / Портал Министерства информационных технологий и связи РФ // http://www.minsvyaz.ru/site.shtml?id=2862.

6. *Гаскаров Д. В.* Интеллектуальные информационные системы. М.: Высш. шк., 2003. 431 с.

7. *Inmon W.* Building the Data Warehouse. N. Y.: John Wiley&Sons, 1992.

8. *Черняк Л.* Эти разные тонкие клиенты // Открытые системы. 2004. № 4. С. 20–30.

A. Yu. Medvedev

CONCEPTUAL BASIS REGARDING COLLECTION OF INITIAL ELECTRONIC DATA

The article considers conceptual basis on creating technology of management and transfer of initial data in terms of intelligent information systems.

---

*) Конечный пользователь — лицо, осуществляющее сбор и ввод первичных данных в систему (подотчетное лицо).