# ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК 025.2:004.91 ВИНИТИ

А. В. Шапкин

# Автоматизированная система комплектования и регистрации входного потока ВИНИТИ. Часть II*

*Дается общая характеристика программного обеспечения автоматизированной системы комплектования и регистрации входного потока ВИНИТИ; рассматриваются модели жизненных циклов объектов обработки; операции, выполняемые над объектами в основном технологическом процессе. Показаны возможности реализации дополнительных функций и включения новых процессов, открывающиеся благодаря принятым архитектурным решениям.*

## 2. УПРАВЛЕНИЕ ЖИЗНЕННЫМИ ЦИКЛАМИ ОБЪЕКТОВ ОБРАБОТКИ

В первой части данной статьи (см. [1]) было рассмотрено информационное поле Автоматизированной Системы Комплектования и Регистрации входного потока (АСКР), построенное на реляционной модели и поддерживаемое средствами SQL (Structured Query Language — Язык структурированных запросов).

Информационное поле АСКР образует единое пространство, на котором взаимодействуют все субъекты — комплектаторы, учетчики, регистраторы, библиографы, разметчики, диспетчер, обеспечивая согласованное и управляемое продвижение по технологическому циклу объектов обработки — изданий, экземпляров, выпусков изданий, документов.

### 2.1. Программное обеспечение

АСКР реализована в архитектуре "клиент-сервер". В качестве системы управления базами данных (СУБД) используется Microsoft SQL-Server. Клиентские программы разработаны при помощи систем программирования Delphi и Visual C++.

Для доступа к данным со стороны приложений применяются технологии ODBC (Open Database Connectivity — Открытый интерфейс доступа к базам данных) и ADO (AcviveX Data Objects).

Следует сделать общее замечание: при построении программного комплекса АСКР (так же, как и при проектировании базы данных) разработчиками уделялось серьезное внимание вопросам "очистки" данных и оптимизации производительности. Важность решения этих проблем для успешного функционирования системы многократно подчеркнута многими авторами. Статьи [2, 3] в наиболее сжатом виде характеризуют эти проблемы.

Если на этапе проектирования информационного поля в свете указанных задач были приняты решения о выделении самостоятельных объектов (путем декомпозиции исходного описания публикации), объединении однотипных объектов в классы, хранении описаний объектов в массивах данных с глубокой нормализацией отношений, то со стороны программного комплекса для обеспечения чистоты данных и производительности выделяются специальные средства контроля алфавита, проверки структуры описаний объектов, выявления дублей в массивах, поддержки целостности технологических цепочек обработки объектов.

#### Серверная часть

Первый уровень составляют простые средства контроля данных в таблицах, предоставляемые SQL, — так называемые ограничения (*constraints*). Это ограничения обязательности полей (*not null*), уникальности полей (*primary key, unique key*), ссылочной целостности (*foreign key*), зависимости между значениями полей (*check constraint*). Они активно используются при построении массивов данных и установлении взаимосвязей между ними.

Более сложные алгоритмы оформлены в виде хранимых процедур (*stored procedure*), которых в системе насчитывается более 350. Развитые возможности SQL (в частности, язык Transact-SQL) позволяют с помощью хранимых процедур анализировать данные, выявлять нарушения или логические несоответствия, динамически актуализировать агрегатные элементы, создавать списки объектов по формальным критериям отбора, формировать несложные отчеты. Хранимые процедуры могут выполняться либо автоматически при модификации данных (триггеры, реагирующие на команды *insert, update, delete*), либо в пакетном режиме (по расписанию), либо по командам из клиентских программ.

Две группы процедур, могут представлять самостоятельный интерес.

* Настоящая публикация является продолжением статьи [1].

*Процедуры формирования хэш-ключей объектов по заданным элементам их описаний и определения сложести объектов.* Это оригинальная разработка; она описана ее автором в [4] и названа нечетким поиском по хэш-ключам. В основе метода лежит построение биграмм – пар последовательных символов. Мерой различия двух строк является "относительная написательная близость", вычисляемая как отношение количества биграмм, различающихся в сравниваемых строках, к суммарному количеству биграмм в обеих строках. Изменением параметров вызова процедур можно задавать значение уровня относительной близости, которое критично для исследования.

Метод нечеткого поиска по хэш-ключам применяется для выявления сходства объектов при определении дублей в массивах данных АСКР. Хэш-ключи формируются и записываются в качестве характеристических ключей в агрегатные таблицы соответствующих массивов данных, что автоматически поддерживается триггерами, реагирующими на изменения описаний объектов. Процедуры вычисления написательной близости объектов запускаются в интерактивном режиме (вызов встроен в интерфейс прикладных программ ведения сериальных изданий и организаций и сопряжен с порождением нового объекта) или в отложенном режиме – по расписанию (в этом случае выдается совокупный отчет о сходстве объектов, модифицированных за некоторый период; такая методика используется, в частности, для анализа изданий книжного типа).

*Процедура формирования ключа сортировки* реализует очистку заданной строки от незначащих слов и символов. Набор элементов описания объекта, из которых надо сформировать ключ сортировки, задается во входных параметрах. При обработке используется список стоп-слов и таблица подстановок символов, учитывающая особенности используемого алфавита. Результат записывается в агрегатную таблицу соответствующего массива данных. В АСКР ключи сортировки автоматически поддерживаются триггерами массивов сериальных изданий и организаций.

### Клиентские прикладные программы

Клиентскую часть программного обеспечения удобно поделить на три группы.

1. Базовые средства, предоставляющие общие для всех приложений функции.

*Поддержка алфавита* — библиотека функций, обеспечивающих средства ввода, визуализации и *контроля текстов, содержащих все символы, предусмотренные* для Реферативного журнала. Алфавит **ВИНИТИ** включает русские, латинские, греческие буквы, цифры, знаки препинания, формульные знаки, специфические буквы европейских языков (диакриты, лигатуры); допускаются несложные формулы с использованием верхних и нижних индексов с уровнем вложенности не более двух.

*Сборка описаний объектов* — библиотека функций, обеспечивающих формирование элементов данных (библиографических и технологических) для заданного объекта: сериального издания, выпуска издания, документа. Нужный объект задается своим идентификатором и источником данных (DSN — Data Source Name) – в качестве параметров вызываемой функции. Сборка значений элементов описания объекта производится в соответствии с построением массива данных и его взаимосвязями с другими массивами, как это было рассмотрено в [1].

Средства поддержки алфавита и сборки описаний объектов представляют собой динамически загружаемые библиотеки (DLL). Они независимо устанавливаются на клиентские рабочие места в среде Windows и могут использоваться всеми приложениями. Визуальные средства работы со сложными текстами используются Delphi-приложениями, будучи оформленными как управляющие элементы ActiveX.

Вынесение на базовый уровень средств поддержки алфавита и сборки описаний объектов позволяет развивать алфавит и менять формы представления объектов в соответствии с изменениями внешних требований — без внесения каких-либо исправлений в эксплуатируемые прикладные программы.

2. Специализированные программы-клиенты — функционально ориентированные приложения, предназначенные для выполнения определенными пользователями определенных операций по управлению объектами в процессе их жизненного цикла или по получению информационных услуг:

для работы с сериальными изданиями – "Ведение массива сериальных изданий", "Регистрационные описания", "Списки и отчеты", "Поиск лакун", "Управление корреспонденцией";

для работы с экземплярами и выпусками — "Учет экземпляров", "Регистрация выпусков СИ", "Регистрация выпусков ИКТ", "Регистрация депонированных рукописей", "Печать библиографических карточек", "Библиографическая обработка", "Диспетчер", "Прием на хранение";

для работы с документами — "Регистрация вторичных документов";

для ведения дополнительных массивов "Массив организаций": "Словари и справочники", "Загрузка Каталога поступлений".

Приведенный список не является полным. Он содержит приложения, реализующие основные операции. Помимо них в систему включены утилиты, специализированные для выполнения вспомогательных функций, не имеющих прямого отношения к работе АСКР: загрузка данных из Ulrich's Periodicals Directory, поддержка технологии микрофильмирования, производство указателей депонированных рукописей и научных мероприятий и пр.

Разбиение программного комплекса на множество относительно независимых прикладных программ можно считать недостатком, так как непросто обеспечить единый интерфейс, приходится испытывать трудности при модернизации. С другой стороны, такое построение позволяет легко конфигурировать автоматизированные рабочие места: на компьютерах комплектаторов, учетчиков, регистраторов, библиографов, разметчиков, диспетчера устанавливаются программы, необходимые для выполнения нужных им операций.

3. Программы общего назначения. В эту группу включены клиентские программы, разработанные для АСКР, но применимые более широко, так как они не имеют жесткой зависимости от структур данных. В основном, это средства экспорта — импорта:

конвертор из HTML-формата в формат ISO-2709 с настройкой входных и выходных меток;

программа выгрузки данных из SQL-таблиц в файлы формата ISO-2709 с настройкой правил сборки данных из связанных таблиц и с гибким управлением логической структурой выходной записи.

### 2.2. Жизненный цикл сериального издания

Формализация жизненного цикла сериального издания опирается на систему технологических состояний и граф возможных переходов от одного состояния к другому с указанием на дугах особенностей и условий переходов. В системе предусмотрено 55 технологических состояний и специфицировано более 300 событий, вызывающих смену состояния (мотивов перехода). По мере изменения внешних условий и развития модели эти значения могут возрастать.

Изменение состояния влечет изменение требований к составу элементов данных описания сериального издания: от предварительного — до полного библиографического описания. Списки необходимых титульных данных являются атрибутами технологических состояний.

Ряд переходов из состояния в состояние сопровождается порождением исходящих писем в адрес тех или иных партнеров. Поэтому параметры для автоматического формирования в массиве корреспонденции соответствующих объектов-писем являются атрибутами причин переходов.

Технологическое состояние сериального издания описывается четверкой атрибутов:

<текущее состояние>, <предыдущее состояние>, <мотив перехода>, [<идентификатор объекта-письма>]

История изменения технологических состояний автоматически накапливается в специальной таблице массива сериальных изданий (поддерживается триггерами). Для каждого события фиксируются код сериального издания, результирующее состояние, дата перехода и пользователь, инициировавший переход.

На рис. 1 представлена модель жизненного цикла сериального издания.

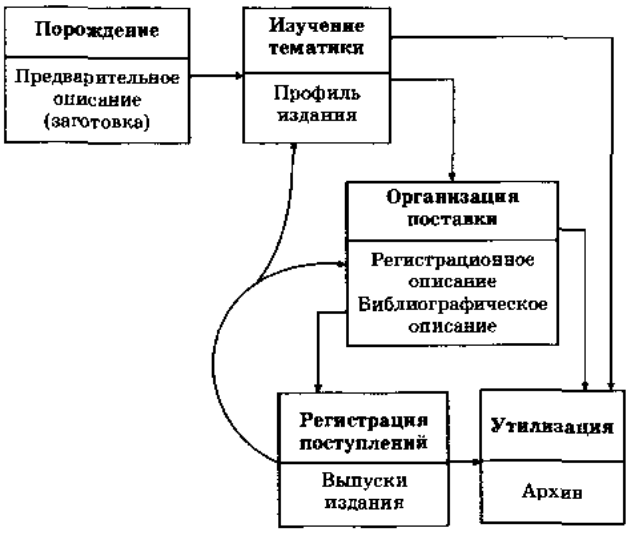


Рис. 1. Модель жизненного цикла сериального издания

**Порождение** сериального издания как объекта обработки в АСКР может быть связано с несколькими причинами. Стандартный путь — ручное заведение комплектатором описания вновь появившегося издания или нового описания взамен старого при изменении названия. Другой источник объектов — загрузка в массив сведений об изданиях из внешних информационных систем.

В любом случае вновь порожденные объекты проходят проверку на дубль (методом сравнения хэш-ключей по сходству) и находятся в технологическом состоянии "предварительное описание" с указанием источника (реклама, подписной каталог, Интернет-сайт, он частного лица, по экземпляру, из базы данных Ulrich's Periodicals Directory и др.). Это состояние предполагает минимальный набор обязательных свойств и требует выполнения дальнейших действий над изданием — в активной фазе.

**Активная фаза жизни** сериального издания включает стадии: 1) изучение тематического профиля, 2) заключение договора о поставке, 3) поступление выпусков во входном потоке.

1. Изучение тематического профиля производится на основе сведений, предоставляемых издателем, оценки экспертов ВИНИТИ, данных об издании из внешних информационных источников. Для проведения этих работ комплектаторы могут запрашивать у издателя образцы издания. Технологические состояния, характерные для данной стадии обработки: "запрошен образец"; "образец получен"; "находится на экспертной оценке"; "требует повторной экспертной оценки"; "вне интересов ВИНИТИ"; "есть тематический профиль". Мотивы изменения состояния и наличие писем, сопутствующих переходам, с очевидностью следуют из приведенных названий.

Результатом первой стадии жизни является вывод о соответствии изучаемого издания тематическому профилю ВИНИТИ и переход к следующей стадии или отказ от дальнейшей обработки (утилизация).

Если издание принято к обработке, то позднее оно может быть возвращено на стадию тематического профилирования для уточнения данных методами вторичной оценки — на основе анализа отражения публикаций в Реферативном журнале (см. подраздел 3.3).

2. Заключение договора о поставке предусматривает согласование с поставщиком наиболее выгодных условий получения выпусков издания. В массиве сериальных изданий достижение этой цели выражается в переводе соответствующих объектов в одно из следующих технологических состояний: "бесплатное получение"; "поставка в обмен на отражение в РЖ (т. н. "зеленая черта")", "поставка по международному книгообмену (МКО)"; "оформлена подписка"; "получение в режиме временного пользования"; "доступ через Интернет".

Решая вопрос об условиях получения выпусков сериального издания, комплектаторы осуществляют многочисленные контакты с потенциальными поставщиками литературы, что отражается в промежуточных технологических состояниях типа "направлен запрос...", "получено согласие...", "получен отказ...", "превышено время ожидания

ответа". Мотивы переходов между этими состояниями понятны из их названий; в нужных случаях переходы сопровождаются автоматическим порождением исходящих писем в адрес конкретных партнеров.

Формальным результатом второй стадии активной фазы жизни сериального издания является так называемое *регистрационное описание* или "договор" поставки (в кавычках потому, что договора в юридическом понимании может и не быть). Регистрационные описания входят в состав технологических данных массива сериальных изданий; они сведены в группу реляционных таблиц. Каждое регистрационное описание действует определенный период времени, связано с конкретным поставщиком (в массиве организаций) и условиями получения выпусков. Другими словами, это отношение вида:

<КСИ>, <период действия>, <код организации-поставщика>, [<условия получения>]

Регистрационные описания периодических изданий в обязательном порядке сопровождаются *сетками ожидания* — формализованными списками выпусков, получение которых ожидается в указанном периоде. Сетки ожидания строятся автоматически — по издательским схемам. В нестандартных случаях сетка может быть исправлена вручную.

Параллельно с организацией поставки производится уточнение титульных данных издания с целью оформления полного *библиографического описания*, а также уточняется *типовой маршрут* обработки выпусков. Типовой маршрут определяется тематикой, важностью издания и зависит от канала поступления.

Таким образом, по завершении второй стадии объект "сериальное издание" оснащен всеми данными, необходимыми для выполнения процедур регистрации выпусков на некоторый период времени (как правило, это календарный год).

3. Регистрация поступающих выпусков — это продуктивная стадия жизни сериального издания, сопровождающаяся порождением в системе новых объектов — выпусков, жизненный цикл которых будет рассматриваться ниже.

Относительно самого сериального издания, находящегося на третьей стадии активной фазы, следует отметить, что система контролирует полноту потока выпусков. Для этого используются автоматизированные средства поиска *лакун*, которые основаны на анализе заполнения сеток ожидания. Обнаружение недостач инициирует изменение технологического состояния, сопровождающееся формированием письма-рекламации в адрес поставщика, и требует от комплектаторов определенных действий по исправлению ситуации.

Кроме того, изменение технологического состояния сериального издания может вызвать диспетчер в случаях отклонения от нормального хода обработки выпусков (временный запрет на регистрацию).

Завершение третьей стадии в жизни сериального издания обусловлено истечением периода действия регистрационного описания (тогда возврат на вторую стадию — в технологическое состояние, предусматривающее продление/изменение договора поставки) или завершением активной фазы (переход к утилизации).

**Утилизация** сериального издания производится при потере к нему интереса со стороны ВИНИТИ или при прекращении его выпуска издателем. В результате утилизации сериальное издание переходит в пассивное состояние, а его описание должно быть помещено в архив.

### Специализированные средства

На автоматизированных рабочих местах комплектаторов и библиографов работа с сериальными изданиями поддерживается, в первую очередь, прикладными программами ведения массива сериальных изданий, управления деловой корреспонденцией и ведения регистрационных описаний. Кроме того, в АСКР создано несколько приложений вспомогательного характера, которые обеспечивают для комплектаторов средства формирования различных списков и отчетов по множеству сериальных изданий, отбор описаний изданий из базы данных Ulrich's Periodicals Directory и загрузку их в массив, средства вторичной оценки изданий по результатам статистической обработки выпущенных информационных продуктов. Для работы этих приложений в структуру массива сериальных изданий включены дополнительные таблицы-отношения (например: таблица связи Ulrich's-АСКР, таблица комплексных отчетов комплектатора).

### Количественная характеристика массива

По состоянию на январь 2005 г. массив сериальных изданий содержит 63952 объекта, описания которых накоплены за период работы АСКР. Динамика пополнения массива представлена в табл. 1:

*Таблица 1*

| Режим эксплуатации АСКР | Год | Прирост массива сериальных изданий | Количество объектов в массиве |
|---|---|---|---|
| Начальная загрузка из Базового массива АЦОП | 1996 | 25743 | **25743** |
| Совместная эксплуатация АЦОП и АСКР | 1997 | 804 | |
| | 1998 | 935 | |
| Опытно-промышленная эксплуатация АСКР | 1999 | 665 | **28147** |
| Промышленная эксплуатация АСКР (новые КСИ >100000) | 2000 | 1367 | |
| | 2001 | 1789 | |
| | 2002 | 27626 | |
| | 2003 | 2518 | |
| | 2004 | 2515 | **63962** |

*Замечания к табл. 1:* а) большой прирост 2002 г. вызван загрузкой описаний журналов из Ulrich's Periodicals Directory: они помещены в массив в технологическом состоянии "в сфере интересов комплектования (предварительное описание)" — для последующего анализа; б) в 2003–2004 гг. начаты работы по размещению в массиве сериальных изданий описаний книжных серий.

С учетом замечаний, можно сделать вывод о том, что среднегодовое пополнение составляет до 1500 объектов-журналов и до 900 книжных серий.

Приводимый в табл. 2 моментальный снимок (от 11 января 2005 г.) дает представление о положении объектов массива сериальных изданий на стадиях жизненного цикла с указанием технологических состояний (даны укрупненные состояния):

*Таблица 2*

| Стадия ЖЦ | Технологическое состояние | Кол-во объектов |
|---|---|---|
| ...после порождения | В сфере интересов комплектования (предварительное описание) | 22403 |
| (1) Изучение тематического профиля | Запрошен образец издания | 2137 |
| | Получен отказ в образце | 152 |
| | Вне интересов ВИНИТИ или не может обрабатываться | 686 |
| | Издание оценено положительно | 1745 |
| | другие состояния | 8 |
| (2) Заключение договора о поставке | Запрос на бесплатную поставку | 549 |
| | Отказ в бесплатной поставке | 9 |
| | Запрошено получение по "зеленой черте" | 1763 |
| | Отказ в "зеленой черте" | 64 |
| | Запрошено получение по МКО | 19 |
| | Предложен МКО (партнером) | 11 |
| | Возврат (неверный адрес) | 82 |
| (3) Регистрация поступающих выпусков | Бесплатное получение | 5924 |
| | Прекращено бесплатное получение | 2 |
| | Поставка по "зеленой черте" | 2917 |
| | Прекращена поставка по "зеленой черте" | 107 |
| | Поставка по МКО | 1096 |
| | Прекращена поставка по МКО | 92 |
| | Издание выписано | 1061 |
| | Поставка в режиме "временное пользование" | 614 |
| | Получение в электронном виде через Интернет (в т. ч. E-library) | 444 |
| | Дезидерата | 5744 |
| | Книжная серия | 1670 |
| | другие состояния активной фазы | 232 |
| Утилизация | Кандидат в архив | 10961 |
| | Архив | 2460 |

Таким образом, в активной фазе находится около 27 тыс. изданий; из них более 19 тыс. — на продуктивной стадии жизни. Эти показатели довольно стабильны.

### 2.3. Жизненный цикл экземпляра и выпуска издания

Моделью жизненного цикла экземпляра является *технологический маршрут* — последовательность операций, выполняемых с экземпляром. Каждая операция может выполняться на одном из специализированных участков. Специализация участков, как правило, проведена по видам литературы и по языкам и странам издания.

В массиве экземпляров маршруты представлены цепочками связанных записей вида:

<УНМ>, <номер операции>, <код операции>, <код участка>, <дата поступления>, <дата начала>, <дата окончания>, <исполнитель>, <код завершения>, <дата передачи на следующий участок>

Для каждого экземпляра, идентифицируемого учетным номером (УНМ), цепочка операций выстроена по возрастанию атрибута <номер операции>. Как видно из набора атрибутов, система позволяет не только планировать обработку, но и фиксировать затраты времени на выполнение операций, а также хранить историю обработки экземпляров.

Безусловно обязательными операциями являются *учет экземпляра, регистрация выпуска, передача экземпляра на хранение*.

**Учет.** Каждый экземпляр научно-технической литературы, поступающий в ВИНИТИ, маркируется наклейкой со штрих-кодом, который содержит учетный номер*. На информационном поле АСКР это событие отражается порождением нового объекта-экземпляра, идентификатором которого служит учетный номер. *После порождения описание экземпляра содержит минимальные данные*: дату учета, способ получения, вид издания (журнал, книга и пр.), тривиальный маршрут (учет—регистрация · утилизация).

**Регистрация выпуска.** Целью регистрации является определение конкретного выпуска издания, представленного тем или иным экземпляром входного потока. В процессе регистрации производится полное библиографическое описание, на основании которого в первую очередь осуществляется проверка на дубль.

Участки регистрации специализированы по видам издания и языкам/странам:

- сериальные издания отечественные
- сериальные издания стран Восточной Европы
- сериальные издания Великобритании
- сериальные издания Германии
- сериальные издания Скандинавских стран
- сериальные издания стран романской группы
- сериальные издания США
- сериальные издания стран Азии и Африки
- сериальные издания, получаемые по валютной подписке
- сериальные издания; срочная регистрация экземпляров временного пользования
- книги иностранные
- книги отечественные
- диссертации и авторефераты (отечественные и иностранные)
- депонированные рукописи
- электронные издания
- издания на компакт-дисках.

В зависимости от вида издания в системе реализованы особые средства для регистрации выпусков сериальных изданий, изданий книжного типа, депонированных рукописей. Специализация выражается в различных методах идентификации изданий, разных требованиях к библиографическому описанию и, соответственно, к алгоритмам определения дублей. Механизмы назначения маршрута обработки выпусков также зависят от вида издания. Отметим некоторые из этих особенностей.

* *Наклейка штрих-кодов допустима только для печатных изданий, являющихся собственностью ВИНИТИ. Особыми случаями являются печатные издания, на которые нельзя клеить штрих-код (личные экземпляры, экземпляры, полученные во временное пользование и т. п.), фиктивные экземпляры (электронные издания, файлы — коллекции документов). Для этих случаев предусмотрены специальные методики учета.*

Регистрация выпусков сериальных изданий базируется на массиве сериальных изданий и сетках ожидания выпусков, поэтому поиск дублей упрощен; для формирования маршрута обработки используется типовой маршрут, хранимый при описании сериального издания. Однако для регистрации выпуска требуется наличие соответствующего объекта-издания, который должен быть безошибочно найден в массиве сериальных изданий и оснащен регистрационным описанием. При отсутствии таких данных для регистрации необходимы контакты со службой комплектования, которая должна ввести соответствующие данные в массив сериальных изданий.

Регистрация изданий книжного типа включает ввод всех необходимых элементов библиографического описания. В связи с этим важно иметь хороший механизм определения дубля, что реализовано методом нечеткого поиска объектов по сходству, *упоминавшимся выше*. *Кроме того, для сокращения* клавиатурного ввода описаний объектов предусмотрен режим регистрации по внешнему источнику, который позволяет использовать готовые библиографические описания. Такой режим использовался, например, при регистрации книг, поступающих из Российской книжной палаты (РКП), когда поток физических экземпляров сопровождался файлом, в котором по коду РКП можно выбирать нужные записи.

Регистрация депонированных рукописей имеет те же особенности, которые отмечены для книг. Дополнительно к этому, при регистрации депонированных рукописей производится их индексирование рубриками ГРНТИ, что можно и нужно использовать для тематической разметки.

Результатом регистрации является одно из двух событий: *фиксация экземпляра-дубля* или назначение экземпляру статуса *"главный"* с порождением нового объекта — выпуска издания — в соответствующем регистрационном массиве.

При фиксации дубля экземпляр, проходящий регистрацию, получает маршрут, который исключает дальнейшую обработку и направляет данный экземпляр на утилизацию*.

Порождение нового выпуска издания (документа монографического уровня) сопровождается выдачей паспорта документа в виде *библиографической карточки* и присвоением *маршрута обработки выпуска*, который включает те или иные технологические операции в зависимости от вида издания, определенного при регистрации.

**Передача на хранение.** Жизненный цикл объекта-экземпляра завершается операцией передачи на хранение (утилизацией) по исчерпании маршрута. Отработанные физические экземпляры направляются по месту хранения — в научные фонды ВИНИТИ (с формированием каталожных карточек), или возвращаются владельцу, если были получены во временное пользование (личные экземпляры, литература БЕН, ГПНТБ и других библиотек), или используются для книгообмена (специально помеченные экземпляры-дубли).

Сведения о прохождении маршрута (история) сохраняются в системе в течение отрезка времени, необходимого для выполнения возможных действий по анализу функционирования, получению кумулятивных отчетов, разбору нестандартных ситуаций. Обычно период хранения истории не составляет больше двух лет. По истечении этого срока технологические данные должны архивироваться.

### Маршрут обработки выпуска (главного экземпляра)

Для главного экземпляра маршрут наполняется конкретными операциями обработки выпуска с указанием технологических участков.

*Печать формуляра* — необязательная операция; выполняется только для публикаций, поступивших в электронной форме.

*Первый контур* — это условное название обобщенной необязательной операции, которая выполняется для специально помеченных изданий. В частности, может включать обработку на участках: Сканирование печатного выпуска (СКАН). Копирование оглавления выпуска — для оперативного информирования ведущих ученых (ГА).

*Разметка* — обязательная операция. Выполняется разметчиком, который просматривает выпуск издания и а) определяет соответствие каждой публикации тематическому профилю редакций и ставит соответствующие штампы тематической разметки 1-го уровня, б) оформляет элементы описания публикации на аналитическом уровне, в) переводит заглавия иностранных публикаций на русский язык (кроме английских, немецких, французских). Разметка выполняется на специализированных участках (по видам издания и языкам/странам — тот же состав, что и вышеприведенный для регистрации). Литература, получаемая во временное пользование, размечается на участке срочной разметки.

*Библиографический контроль* — обязательная операция. Выполняется библиографом, который контролирует правильность описания выпуска на монографическом уровне, правильность библиографической разметки на аналитическом уровне, а также при необходимости выделяет разделы в выпуске (материалы научных мероприятий).

*Ввод описаний статей* — необязательная операция, выполняемая для изданий, которым назначена полная аналитическая обработка.

*Ксерокопирование* — операция обязательно выполняется для печатных изданий. Заключается в создании копий первичных документов (с технологическим паспортом в виде макетированной страницы) и рассылке их в научные редакции в соответствии с кодами тематической разметки. Наряду с основным участком копирования предусмотрен участок срочного копирования, на который направляются экземпляры временного пользования.

*Микрофильмирование* — необязательная операция. Выполняется для "чужих" печатных выпусков (временное пользование, личные экземпляры). Поддерживается средствами Автоматизированной подсистемы микрофильмирования (АСФИМ)**.

---

* При развитии системы возможно реализовать параллельную обработку экземпляров-дублей по разным маршрутам.

** Процессы микрофильмирования фиксируются в массиве данных АСФИМ и обеспечиваются специальными приложениями. Как массивы данных, так и программные средства АСФИМ органично включены в архитектуру АСКР. На их основе осуществляется взаимодействие служб комплектования, регистрации и микрофильмирования.

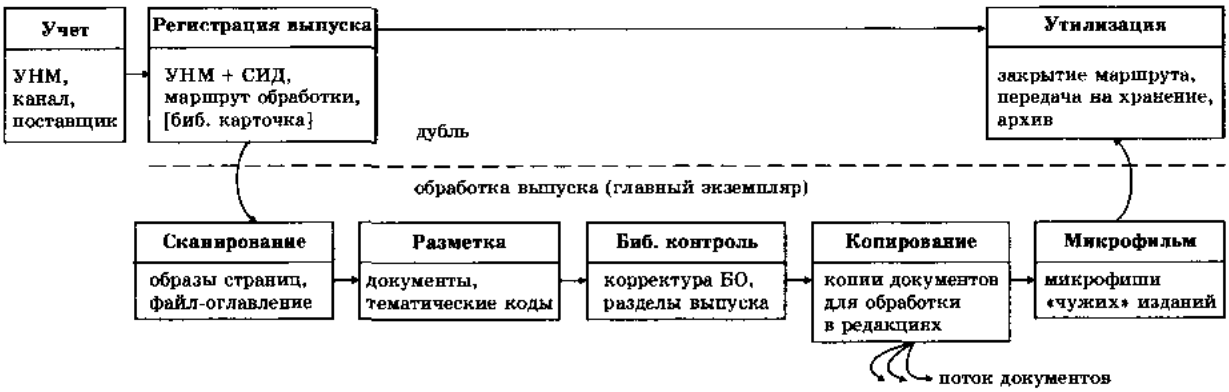


Рис. 2. Примеры технологических маршрутов главного экземпляра и дубля

*Второй контур* — это условное название обобщенной необязательной операции, которая выполняется для специально помеченных изданий. В частности, может включать обработку на участках: Рассылка копий статей по заявкам ведущих ученых (ГА-2), Подготовка экспресс-информации по физике (ФИЗ), Сигнальная информация по российским журналам (СИГНАЛ).

*Диспетчеризация* — особая операция, служащая для направления экземпляра диспетчеру. Используется для фиксации в массивах АСКР продвижения физических экземпляров по участкам в результате выполнения ручных операций (тех, которые не имеют непосредственной связи с системой), а также в целях разбора нестандартных ситуаций, обнаруженных при обработке выпуска.

Набор и последовательность операций зависит от вида издания, канала поступления и других факторов, определяемых диспетчером в нужных случаях. На рис. 2 показаны примеры технологических маршрутов.

### Количественная характеристика массивов экземпляров и выпусков

Динамика пополнения массивов экземпляров и выпусков с начала эксплуатации АСКР приведена в табл. 3:

*Таблица 3*

| Год | Кол-во учтенных экземпляров | Кол-во зарегистрированных выпусков изданий |
|---|---|---|
| 1999 | 98008 | 87953 |
| 2000 | 113515 | 99922 |
| 2001 | 100784 | 82947 |
| 2002 | 120778 | 102582 |
| 2003 | 129452 | 111003 |
| 2004 | 107201 | 94686 |

Относительное превышение объема потоков в 2002 и 2003 гг. объясняется форсированным "разгребанием завалов" необработанной литературы, скопившейся до начала эксплуатации АСКР. Установившийся режим характеризуется такими показателями: в год на обработку поступает примерно 120 тыс. экземпляров НТЛ, при обработке которых регистрируется до 95 тыс. выпусков изданий всех типов.

По среднестатистическим данным, в системе на обработке ежемоментно находится примерно 35 тыс. выпусков изданий (главных экземпляров). Моментальный снимок от 11 января 2005 г. (табл. 4) дает представление о нагрузке по конкретным технологическим операциям:

*Таблица 4*

| Технологическая операция | Кол-во ожидающих экземпляров |
|---|---|
| Регистрация выпуска | 9079 |
| Печать формуляра | 256 |
| Первый контур | 141 |
| Разметка | **12852** |
| Библиографический контроль | 4811 |
| Ввод описаний статей | 44 |
| Ксерокопирование | 4366 |
| Микрофильмирование | 179 |
| Второй контур | 542 |
| Диспетчеризация | 1393 |
| Передача на хранение | 129 |

Отсюда, в частности, видно, что самой трудоемкой является операция разметки. Наиболее загружен участок разметки литературы стран Азии и Африки (ожидают разметки 3481 экз.).

## 2.4. Аналитическая обработка выпусков

Аналитическая обработка предполагает исследование содержания выпусков и порождение новых объектов. В АСКР различаются два уровня аналитической обработки выпусков изданий:

1) выделение в выпуске разделов — практически применяется для создания обобщенных описаний материалов конференций, симпозиумов, съездов и других научных мероприятий; соответствующий массив данных был описан в [1, подраздел 1.3];

2) выделение "несамостоятельных" публикаций (статей в журналах, книжных сборниках и т. п.) и их последующая обработка — этот процесс рассматривается в данном подразделе как естественное продолжение обработки выпусков изданий.

На информационном поле АСКР результатом аналитической обработки является порождение документов (с привязкой к родительским выпускам изданий) и направление их на содержательную обработку в соответствии с тематическими аспектами. Для выполнения этих операций предусмотрен *массив документов*, непосредственно связанный с

Каталогом поступлений, и разработан функционально полный инструментарий для манипулирования документами: идентификация, библиографическое описание, выявление дублей, редактирование, тематическая разметка, выдача паспорта и пр.

**Порождение документа** предусматривает заведение библиографического описания публикации, присвоение документу уникального идентификатора и выдачу паспорта публикации.

*Библиографическое описание* — первое действие, выполняемое при регистрации документа. Важной задачей является выявление дублей публикаций во входном потоке и установление преграды для порождения документов-дублей в технологической базе данных. Решению этой задачи помогают механизмы динамического сравнения документов по характеристическим ключам.

*Паспорт документа.* Порождение документа завершается выдачей твердой копии паспорта публикации, который представляет собой формуляр с воспроизведенными на нем идентификатором и библиографическим описанием.

**Активная фаза.** В активной фазе жизни документ последовательно проходит ряд состояний, что отражает реальный процесс выполнения операций над публикацией. Формуляр сопровождает публикацию по всем участкам обработки: в него заносятся элементы содержательного описания, на нем делаются пометки о выполнении технологических операций.

В АСКР над документом может выполняться одна операция — разметка 1-го уровня. Это определение тематики публикации с точностью до редакции. В общем случае может быть выделено несколько тематических аспектов.

Таким образом, *аналитическая обработка обеспечивает сырьем (документами) дальнейший производственный процесс.* При этом важным фактором является наличие у документов тематических кодов, которые являются "адресами" научных редакций, ответственных за содержательную обработку. На этой стадии управление жизненным циклом документов выходит за рамки обработки входного потока и перемещается в другие компоненты технологической базы данных.

**Количественная характеристика.** По состоянию на январь 2005 г. массив документов содержит 250 тыс. объектов. Источниками являются:

| | |
|---|---|
| ручной ввод единичных описаний документов | 1658 шт. |
| загрузка оглавлений журналов после сканирования | 113 343 шт. |
| загрузка описаний документов, полученных из электронных источников | 103 366 шт. |
| загрузка описаний документов из базы данных русскоязычных журналов | 36 429 шт. |

Динамику пополнения массива можно представить в следующем виде:

| | | |
|---|---|---|
| до 2000 г. | 44 516 шт. | Основной источник — оглавления журналов после сканирования печатных выпусков |
| 2000 г. | 21 326 шт. | |
| 2001 г. | 18 273 шт. | |
| 2002 г. | 15 643 шт. | |
| 2003 г. | 30 681 шт. | Основной источник — документы после "электронной" регистрации выпусков журналов |
| 2004 г. | 120 918 шт. | |
| 2002 г. | 3439 шт. | |

Видно, что в настоящее время главную составляющую потока документов дает обработка выпусков журналов из электронных источников. В будущем ожидается значительный рост этого потока. В связи с этим в АСКР активно развиваются соответствующие технологии, программное обеспечение и средства управления.

## 3. НОВЫЕ ВОЗМОЖНОСТИ АСКР

Если выше были рассмотрены стандартные функции системы, реализованные в моделях жизненных циклов основных объектов, то теперь хотелось бы обратить внимание на новые возможности АСКР, обеспечиваемые ее архитектурой.

### 3.1. Обработка публикаций из "электронных" источников

В последние годы наблюдается сокращение доли печатных изданий и возрастание "электронной" составляющей входного потока. Это версии печатных журналов на компакт-дисках; издания, предоставляемые электронными библиотеками в Интернет; материалы научных мероприятий на компакт-дисках; файлы-оглавления журналов и сборников — от издательств (могут сопровождаться и файлами с полным текстом); тематические подборки описаний статей, подготавливаемые научными сотрудниками ВИНИТИ.

Тривиальный метод включения таких публикаций в рабочий поток — распечатка и запуск по обычной технологической цепочке. Отрицательные качества этого способа очевидны: расход бумаги, ручные операции при регистрации изданий, необходимость последующего ручного ввода элементов описаний документов.

Поэтому актуален вопрос обеспечения автоматизированной массовой обработки электронных изданий непосредственно с исходных файлов, а также предоставления возможности дальнейшей обработки документов из таких изданий с включением в конечные информационные продукты — без ручного ввода элементов данных, — за счет полного использования информации, содержащейся в исходных файлах. Основные задачи при этом следующие:

определение метода учета (куда наклеивать учетный номер?);

автоматизированное распознавание изданий на монографическом уровне по описаниям отдельных статей;

непротиворечивая регистрация "электронных" и печатных выпусков (определение дублей, обеспечение правильной обработки в случае получения печатного экземпляра после электронного);

наиболее полное использование всех имеющихся элементов описания публикации, включая авторскую аннотацию и ключевые слова, — без дополнительного ручного набора библиографических данных;

обеспечение возможности дальнейшего продвижения имеющихся описаний публикаций (на аналитическом уровне) в информационные продукты.

При решении поставленных задач необходимо учитывать такие отличительные особенности исходных данных, как *разноформатность* (файлы с описаниями документов могут иметь различный

физический формат и логическую структуру записей) и *неполнота* (не все статьи выпуска издания могут присутствовать в одном файле, они могут быть рассыпаны произвольным образом, могут повторяться и т. д.). Принципиальные решения:

любые входные файлы электронных изданий рассматриваются как коллекции разрозненных описаний публикаций: каждая запись содержит элементы библиографического описания монографического и аналитического уровней, записи обрабатываются независимо друг от друга, полнота выпусков изданий не контролируется;

форматы обработанных файлов специфицируются и накапливаются в системе для возможного повторного использования (обучение по входным форматам);

при учете и регистрации электронных изданий используется понятие "фиктивный экземпляр" с соответствующими соглашениями о технологическом маршруте и приоритетности обработки;

автоматическое распознавание издания и выпуска (по основным идентифицирующим признакам) совмещено с ручным режимом при гибком технологическом взаимодействии между комплектаторами, регистраторами и библиографами.

Указанные принципы были положены в основу построения модели данных, технологии и программного обеспечения учета, регистрации и библиографической обработки публикаций из электронных источников [5].

**Массив данных** регистрации выпусков по электронным документам включает головную транзитную таблицу и связанные с ней таблицы форматов и таблицы-словари.

*Транзитная таблица* содержит описания документов в унифицированной форме. Каждая запись транзитной таблицы имеет статус — индикатор состояния документа. Продвижение документа по технологической цепочке отражается в изменении значения статуса. Записи в транзитной таблице существуют не постоянно, а до тех пор, пока они не загружены в массив документов. После этого они удаляются из транзитной таблицы, либо хранятся там ограниченное время для предотвращения повторной загрузки.

*Таблицы входных форматов* содержат сведения о физических форматах обработанных файлов, а также списки меток элементов данных и их соответствие внутрисистемным меткам АСКР. Таблицы входных форматов пополняются при эксплуатации системы.

Системой поддерживается список загруженных файлов — для анализа результатов работы и предотвращения повторной загрузки.

**Технология.** В настоящее время разработана и широко используется технология автоматизированной обработки файлов, содержащих описания статей из журналов. Формирование входного файла как таковое считается внешней по отношению с АСКР задачей. Например, источниками могут быть Интернет-сайты издательств или электронных библиотек, базы данных. Основные технологические операции, выполняемые в АСКР:

преобразование исходного файла для загрузки в соответствии со специфицированным форматом и приведение текстов полей к алфавиту ВИНИТИ;

загрузка описаний статей в транзитную таблицу — с контролем логической структуры записей;

распознавание сериальных изданий по монографическим элементам; при необходимости, заведение в массиве новых изданий;

регистрации выпусков изданий — определение выпусков по регистрационным массивам или регистрация новых выпусков; помещение новых выпусков в Каталог поступлений с пометной "электронный источник";

помещение описаний статей в массив документов — с контролем дублей;

обработка документов и предоставление сформированных библиографических описаний для информационных продуктов.

Оперативная обработка файлов электронных изданий предъявляет повышенные требования к согласованному функционированию работников на разных участках: администратора-диспетчера, регистраторов, комплектаторов, библиографов. В связи с этим в системе особо регламентированы: *а) оформление и передача файлов электронных изданий*; б) состав ситуаций, возникающих при обработке файлов на этапе загрузки в транзитную таблицу, и порядок взаимодействия с источником файлов; в) состав ситуаций, возникающих при регистрации выпусков журналов, и порядок взаимодействия между регистраторами и комплектаторами; г) оформление результатов обработки входных файлов; д) паспортизация документов и направление их на обработку в научные редакции.

**Программное обеспечение.** Специально для обработки файлов электронных изданий программный комплекс АСКР пополнен двумя крупными приложениями.

"Конвертор HTML-файлов в формат ISO". Решает задачи преобразования страниц, полученных из Интернет, к формату, допускающему загрузку в транзитную таблицу электронной регистрации; предусмотрены настройки по входным и выходным меткам, а также настройки правил подстановок символов для обеспечения правильности алфавита — через таблицы-словари нормативно-справочной системы.

"Регистрация выпусков СИ по электронным документам". Решает задачи: а) загрузка в транзитную таблицу данных из внешних файлов нескольких фиксированных форматов; б) идентификация журналов-источников по массиву сериальных изданий — с обучением посредством накопления прецедентов; в) учет фиктивного экземпляра и регистрация выпуска журнала-источника — с формированием регистрационного описания издания и маршрута фиктивного экземпляра в зависимости от канала поступления исходного файла; г) очистка транзитной таблицы.

Кроме того, существенно доработана программа "регистрация и разметка вторичных документов": реализованы функции загрузки в массив документов из транзитной таблицы и печати паспорта документа, полученного по каналу электронной регистрации.

**Эксплуатация.** Экспериментальная обработка электронных изданий начата в 2003 г., когда опробовались различные форматы данных и варианты технологии. С начала 2004 г. работа была переведена в производственный режим, с постепенным наращиванием входного потока: к середине года на обработку ежемесячно поступают по 10 тыс. описаний журнальных статей (журналы 472 наименований, получение которых в печатном виде

стало невозможно). Основным источником данных этого потока служит Интернет-сайт электронной библиотеки Российского фонда фундаментальных исследований. Помимо основного потока на обработку поступают данные из других электронных библиотек и баз данных, подготовленные специалистами научных редакций ВИНИТИ.

Параллельно продолжаются исследования возможностей расширения комплектования входного потока на основе электронных источников. Главное внимание уделяется крупным издательствам и электронным библиотекам, которые могут предоставить большие коллекции аннотированных публикаций из периодических и продолжающихся изданий по профилю ВИНИТИ.

### 3.2. Научные мероприятия

"Научное мероприятие" — обобщающий термин, принятый для обозначения таких событий, как конференции, конгрессы, съезды, симпозиумы, школы-семинары и тому подобные мероприятия, состоявшиеся или планируемые в научном мире.

**Массив научных мероприятий.** Объектом массива является *описание научного мероприятия*. Характерными (и обязательными) свойствами этих объектов являются: наименование; место проведения — страна, название населенного пункта; время проведения — год, даты; тематика; организаторы; статус (состоится в будущем — прошло — перенесено - отменено). Из этого набора свойств, в частности, следует, что виртуальные конференции не входят в поле зрения; рассматриваются только "реальные" мероприятия.

**Нормализация описания научного мероприятия.** С целью систематизации множества научных мероприятий и для упорядочения структуры описания этих объектов в АСКР реализуется *классификация*, которая учитывает такие признаки, как регулярность мероприятий (например, ежегодная конференция), подчиненность (например, семинар в рамках...), совместное проведение независимых мероприятий, географический охват (международное, национальное, региональное), участники (молодежная конференция,...). Конкретные значения выделенных признаков требуют наличия в описании тех или иных титульных элементов и технологических данных.

Более детальное описание мероприятия может включать развернутую характеристику тематики, программу, перевод титульных данных на русский язык.

Важным свойством научного мероприятия (как объекта в информационной среде АСКР) являются *сведения о материалах*. Предусматривается несколько уровней представления и хранения таких сведений:

а) библиографическая ссылка на материалы, обнаруженные во входном потоке (сборник трудов, тезисы докладов в журнале) – практически это указание на соответствующий раздел зарегистрированного выпуска издания;

б) список описаний докладов на аналитическом уровне (как правило, с авторскими аннотациями) — эти данные хранятся непосредственно в массиве документов АСКР;

в) ссылка на внешний ресурс — Интернет-сайт, на котором представлены материалы мероприятия.

Вариант представления сведений о конкретном мероприятии зависит от доступности соответствующих данных, важности мероприятия, трудоемкости ввода.

Таким образом, описания научных мероприятий связаны с объектами массивов регистрации выпусков изданий, организаций, документов.

**Формирование и ведение массива мероприятий.** Входной поток содержит сведения о научных мероприятиях, поступающие из двух источников.

Во-первых, ВИНИТИ как национальный центр НТИ получает от своих партнеров анонсы предстоящих мероприятий. Их описания заносятся в массив, и комплектаторы осуществляют дальнейший процесс обработки: контакты с организатором, слежение за сроками проведения, уточнение титульных данных и, наконец, получение опубликованных материалов для отражения в информационных продуктах.

Во-вторых, поступающая на обработку научно-техническая литература содержит материалы состоявшихся мероприятий — в журналах, в сборниках докладов, в сборниках депонированных трудов конференций и др. После регистрации таких изданий эти материалы, как правило, выделяются в самостоятельные объекты — разделы соответствующих выпусков изданий, которые являются источником данных для массива мероприятий.

Таким образом, описания мероприятий вводятся *в массив как вручную — по анонсам, так и* посредством "автоматизированной добычи" — при переработке регистрационных данных. Особенностью анонса является приблизительность описания мероприятия, которое впоследствии может уточняться. Особенностями описаний, извлеченных из регистрационных данных, являются возможная многовариантность (когда материалы одного мероприятия опубликованы в различных независимых изданиях, обработанных на входе) и частичность (когда материалы одного мероприятия поделены на несколько томов).

В этих условиях чистоту и целостность массива мероприятий (исключение дублей, правильное связывание регулярных, подчиненность) помогают обеспечивать человеко-машинные механизмы, основывающиеся на:

классификации объектов и связанной с ней структуризации описаний;

поддержке нормальных форм дат, географических названий, специфических терминов — при помощи соответствующих словарей и хранимых процедур;

формировании и использовании характеристических ключей.

Обработка материалов состоявшихся научных мероприятий (вторая составляющая потока) является стандартной операцией, которая выполняется библиографами и регистраторами для журналов и книг-сборников. Результаты этой работы оформляются в виде специальных объектов — разделов (частей) выпусков изданий, содержащих материалы соответствующих конференций, съездов, симпозиумов и пр. (см. [1, подраздел 1.3] и [6]). По состоянию на январь 2005 г. объем массива превысил 17 тыс. записей. Эти данные составляют основу для формирования в Каталоге поступлений самостоятельного указателя научных мероприятий. Он

представлен на Интернет-сайте ВИНИТИ и доступен пользователям в свободном режиме.

Обработка анонсов и полнофункциональное ведение массива мероприятий с приемом машиночитаемых описаний из различных источников являются предметом развития системы, запланированного на ближайшее время.

### 3.3. Исследование свойств объектов на основе анализа отражения публикаций в информационных продуктах

Интересными вопросами являются:

выявление/уточнение тематической направленности периодических и продолжающихся изданий;

определение продуктивности периодических изданий для ВИНИТИ;

связывание классификационных индексов публикаций с Рубрикатором ВИНИТИ (для литературы некоторых видов);

сроки обработки объектов в зависимости от их особенностей.

Источником данных для проведения подобных исследований могут служить информационные продукты ВИНИТИ – выпуски Реферативного журнала, которые в машиночитаемом виде представлены в соответствующих базах данных и доступны для автоматизированной обработки.

Накопление исходных данных производится в специализированной **подсистеме статистической обработки,** которая базируется на информационном поле АСКР. Ее основу составляет **массив образов документов,** связанный с массивом сериальных изданий, массивом организаций и Каталогом поступлений. Каждый документ, загружаемый из базы данных (выпуска Реферативного журнала), представлен в виде сокращенного описания. Образ документа включает наиболее информативные для анализа элементы:

идентификатор выпуска издания (СИД), которому принадлежит публикация;

вид публикации (статья в журнале, книга, депонированная рукопись и пр.);

язык документа;

код Реферативного журнала, в котором отражен документ, год и номер выпуска;

список рубрик, к которым отнесен документ — по Рубрикатору ВИНИТИ;

список ключевых слов и словосочетаний, использованных редактором при координатном индексировании публикации;

значения специфических элементов (номера специальностей для диссертаций, коды классификации стандартов, коды Международной патентной классификации для патентов);

признак заимствования описания документа из других продуктов ВИНИТИ.

Наличие перечисленных элементов позволяет: 1) строго соотносить каждый документ какому-либо выпуску издания — через СИД по Каталогу поступлений; 2) определять тематическую направленность документа с разной степенью точности – через код информационного продукта (выпуск РЖ – это второй уровень тематической разметки, углубляющий первичный код научной редакции), или, еще более глубоко, — через рубрикационные шифры; 3) определять срок завершения обработки документа в ВИНИТИ – через нумерацию выпуска РЖ. Имея выборку образов документов, связанных с Каталогом поступлений и с массивом сериальных изданий, можно ставить задачи проверки правильности тематических кодов журналов, определения весомости тех или иных изданий в тех или иных научных направлениях, контроля сроков обработки (в том числе в зависимости от языка публикации и редакции ВИНИТИ, в которой проводится содержательная обработка). Очевидно, что достоверность получаемых ответов будет определяться представительностью исходной выборки.

**Технология обработки** включает этапы: загрузка данных в массив образов документов и очистка от формальных ошибок, формирование рабочих таблиц, выдача отчетов.

1. Загрузка массива образов документов и очистка данных

На момент написания данной статьи загружены описания публикаций, которые нашли отражение во всех выпусках Реферативного журнала ВИНИТИ за три года. В результате в массив поступило 2 914 064 образов документов (включая заимствованные документы — дубли описаний публикаций, отраженные в разных выпусках РЖ). Распределение по видам публикаций приведено в табл. 5.

*Таблица 5*

| Вид публикации, отраженной в РЖ | Кол-во за 3 года |
|---|---|
| Статья в сериальном издании | 2 132 055 |
| Отдельный выпуск журнала | 1375 |
| Статья в книге, сборнике | 368 055 |
| Книга | 54 849 |
| Проспект, каталог | 476 |
| Нормативно-технический документ | 2185 |
| Диссертационная работа | 32 098 |
| Картографическое издание | 389 |
| Статья в сборнике депонированных рукописей | 1328 |
| Депонированная рукопись | 10 288 |
| Патентный документ | 310 966 |

Элементы каждого описания документа при загрузке подвергаются контролю с привлечением нормативно-справочной информации и массивов основных объектов АСКР. Например, в 7460 документах выявлены ошибки в кодировке языка текста или страны издания. Основная масса таких ошибок поддается исправлению при последующей очистке, что позволяет не исключать документы из статистических расчетов.

Для проведения адекватной оценки продуктивности и тематической направленности изданий наиболее критичной является возможность нарушения структурной целостности цепочки *<Массив сериальных изданий — Каталог поступлений — Образы публикаций>*. К таким ошибкам относятся следующие:

описание публикации не имеет ссылки на выпуск издания (отсутствует СИД);

описание публикации имеет ошибочный СИД, отсутствующий в Каталоге поступлений;

описание статьи не имеет ссылки на сериальное издание (отсутствует КСИ — код сериального издания);

описание статьи имеет ошибочный КСИ, отсутствующий в массиве сериальных изданий;

описание статьи имеет допустимые КСИ и СИД, но они не соответствуют друг другу по Каталогу поступлений.

Общий вес таких дефектных записей превышает 13%. Указанные дефекты возникают, в первую очередь, из-за ошибок набора данных и несовершенства средств формально-логического контроля при производстве Реферативного журнала. Другой источник ошибок — недостатки существующей технологии производства информационных продуктов, которая не учитывает динамики объектов в АСКР.

Для поддержания представительности выборки и повышения достоверности исследований многие из перечисленных дефектов могут быть исправлены при помощи специальных процедур корректировки пар *<СИД КСИ>* в образах публикаций. Методика корректировки опирается на текущее состояние массива сериальных изданий и Каталога поступлений с учетом изменения состояний объектов (слияние-разделение сериальных изданий, замещение регистрационных данных и пр.). В результате применения процедур очистки и нормализации описаний публикаций количество дефектных записей снижается более чем на 2/3.

Следует подчеркнуть важность корректировки пар *<СИД — КСИ>*. Она позволяет привести все *описания документов (в части принадлежности их* к конкретным изданиям) к состоянию на один момент — момент съема сведений с Каталога поступлений и массива сериальных изданий. В этом заключается принципиальное отличие описываемой методики от методов, основанных на прямом анализе множества документов, накапливаемого в банке данных.

2. Формирование промежуточных таблиц

В результате начальной загрузки и формальной очистки получен массив для анализа фактической продуктивности и тематической направленности сериальных изданий, в котором представлены статьи из 221 863 выпусков 12 684 изданий.

Следующим шагом является создание рабочей таблицы, в которой для сериальных изданий агрегированы сведения о количестве публикаций — по кодам тематической разметки 1-го уровня (штампы научных редакций определяются по именам файлов баз данных через Регистр информационных продуктов ВИНИТИ) с разбивкой по выпускам и годам обработки. Ключевая комбинация атрибутов такой таблицы состоит из четырех элементов: *<КСИ, СИД, отдел, год>*. Содержательная часть каждой записи включает общее количество описаний публикаций и количество основных документов (без заимствованных). Для примера в табл. 6 приведен фрагмент, в котором собраны сведения об отражении редакциями ВИНИТИ публикаций из четырех выпусков журнала "Автометрия", издаваемого Институтом автоматики и электрометрии Сибирского отделения РАН. (КСИ=19).

3. Выдача отчетов

Полученная сводная таблица фактического отражения в РЖ статей из сериальных изданий составляет основу для выдачи различных отчетов. Прикладные программы формирования отчетов позволяют варьировать методы агрегирования данных; настраивать фильтры по группам стран, видам и статусам изданий; задавать включаемые в отчет параметры; определять формы выдачи (файл, распечатка) и пр.

*Таблица 6*

| КСИ | СИД | *Отдел (научная* редакция) | *Год* | *Кол-во* докумен-тов в РЖ | *Кол-во* статей |
|---|---|---|---|---|---|
| ... | ..... | .... | .. | ..... | .... |
| **19** | J01746992 | Автоматика и радиоэлектроника | 02 | 6 | 6 |
| **19** | J01746992 | Биология | 02 | 1 | 1 |
| **19** | J01746992 | Математика | 02 | 5 | 5 |
| **19** | J01890058 | Автоматика и радиоэлектроника | 02 | 4 | 3 |
| **19** | J01890058 | Математика | 02 | 3 | 3 |
| **19** | J01890058 | Физика | 02 | 1 | 1 |
| **19** | J02197674 | Автоматика и радиоэлектроника | 02 | 1 | 1 |
| **19** | J02197674 | Физика | 02 | 1 | 1 |
| **19** | J02197674 | Электротехника | 02 | 1 | 1 |
| **19** | J02253388 | Автоматика и радиоэлектроника | 02 | 2 | 2 |
| ***19*** | *J02253388* | *Физика* | *02* | *4* | *4* |
| **19** | J02253388 | Химия | 02 | 1 | 1 |
| ... | ..... | .... | .. | ..... | .... |

Система отчетов ориентирована на аналитическую работу комплектаторов по вопросам определения тематической направленности и продуктивности изданий, выявления тематических профилей организаций-партнеров.

**Использование.** Подсистема статистической обработки, связанная с массивами данных АСКР, предоставила, в первую очередь, хороший полигон для исследований в области оптимизации входного потока информационного центра. Вот два примера таких работ.

*База данных основных периодических изданий.* Подготовлена отделом комплектования на основе описаний изданий из Ulrich's Periodicals Directory. При ее построении проводилось сопоставление описаний изданий со сведениями, хранящимися в массиве сериальных изданий, и результатами статистической обработки сведений об отражении публикаций в РЖ. Используется при комплектовании входного потока иностранных периодических изданий, а также является самостоятельным продуктом, публикуемым на Интернет-сайте ВИНИТИ.

*Перечень основных российских научных и научно-технических журналов.* Построен на основе обработки статистических данных об отражении в РЖ статей из российских журналов. Используется для формирования списка изданий, которые в обязательном порядке должны быть представлены во входном потоке полными комплектами и исчерпывающе отражаться в информационных продуктах ВИНИТИ.

Подробное описание указанных работ можно найти в [7, 8, 9].

Возможности подсистемы статистической обработки не ограничиваются анализом сериальных изданий для целей комплектования. Другие аспекты ее использования будут рассмотрены далее.

### 3.4. Автоматизация тематической разметки

Задача раскладки поступающих публикаций по "тематическим полочкам" для последующего направления их в научные редакции – одна из важнейших в технологической цепочке обработки литературы в информационном центре. Решение этой задачи основывается на многоуровневой системе индексирования: от грубого определения отрасли знания (физика, химия, география,...) до выявления тонких аспектов в соответствии с Рубрикатором ВИНИТИ. Уровни иерархии связаны с организационной структурой: тематическая разметка по научным редакциям образует 1-й уровень, далее следует углубление на 2-й уровень — по конкретным редакторам. В отношении выпускаемых информационных продуктов можно видеть три уровня индексирования: 1-й соответствует сводному тому Реферативного журнала (или нескольким сводным томам, выпускаемых отделом), 2-й — тематическому выпуску РЖ, 3-й уровень – разделу в выпуске РЖ – по рубрикации.

В общем процессе индексирования публикаций задача АСКР — это только разметка на 1-м уровне — по 16-ти адресам, соответствующим научным редакциям (см. [1], рис. 1). При этом всегда выделяется один основной отдел, в который материал направляется на полную обработку с реферированием, и возможные дополнительные отделы, которые используют результаты обработки в основном отделе и добавляют свои классификационные индексы. Таким образом, 1-й уровень разметки задает начальное направление обработки публикаций и является весьма ответственным.

Определим нагрузку на участок разметки. Годовой поток, составляющий 95 тыс. выпусков литературы всех видов (кроме патентов), дает примерно 620 тыс. отдельных публикаций. Нетрудно подсчитать, что, если в штате имеется 10 разметчиков, то каждому из них в день надо обработать 248 публикаций, или более 35 шт. в час*. Значит, пролистывая выпуск издания, разметчик имеет около 2-х минут на осмысление нужности каждой публикации, библиографическую разметку, определение тематики и простановку штампов редакций, соответствующих тематике. И все это при том, что обрабатывается разноязычная литература, и необходимо тратить время на перевод, чтобы хоть примерно понять содержание.

Приведенный грубый расчет позволяет утверждать, что участок разметки является одним из узких мест в общей технологической цепочке. Это видно и из моментального снимка распределения экземпляров по технологическим операциям (см. выше в подразделе 2.3). Конечно, организация работы группы разметки предусматривает предварительную сортировку по языкам, наличие примерных тематических шаблонов для наиболее важных журналов, привлечение специалистов из научных отделов, другие средства, позволяющие повысить производительность. Тем не менее, должно быть понятно, что разработка и внедрение средств автоматизации в этой сфере является весьма благодарной задачей. Особенно, если учесть дефицит квалифицированных специалистов, какими должны быть разметчики.

Автоматизация тематической разметки публикаций в рамках АСКР может опираться лишь на те данные, которые накоплены в информационных массивах системы на момент выполнения обозначенного действия. При этом исключается ввод каких-либо элементов специально для целей разметки. Поэтому будем рассматривать задачу **добычи данных о тематике публикации на основе библиографического описания:** необходимо сделать предположение (а возможно, и точное заключение) о тематической направленности публикации, используя при этом только те сведения, которые имеются в библиографическом описании.

Ставить и решать эту задачу можно, так как библиографическое описание содержит тематически нагруженные элементы – такие элементы, значения которых дают возможность хотя бы грубо судить о тематике. На монографическом уровне — это (в зависимости от вида публикации) сериальное издание, издательство, коллективный автор, специфические элементы для изданий книжного типа. На аналитическом уровне наиболее информативны авторская аннотация и ключевые слова.

Посмотрим, что можно получить от **монографического описания,** поскольку только оно стопроцентно имеется в системе к моменту разметки. Идея состоит в предварительной подготовке данных — "тематическом профилировании" объектов, которые затем будут использованы в описаниях выпусков изданий (тематический профиль объекта — это список кодов разметки 1-го уровня). Архитектура информационных массивов АСКР позволяет осуществить такие работы в отношении сериальных изданий, организаций, научных мероприятий, а также создать специализированные классификационные схемы для некоторых изданий книжного типа.

*Профилирование сериальных изданий* основывается на статистической обработке массива документов, обработанных в ВИНИТИ за последние три года. В результате выявлено 2825 так называемых "моноразметочных изданий" — таких, которые имеют одинаковый набор кодов разметки для всех статей за весь период наблюдения. Эти тематические коды в первую очередь занесены в описания соответствующих объектов массива сериальных изданий. Примеры тематических профилей (табл. 7).

Дальнейшие работы предполагают исследования более сложных случаев разметки журналов, а также углубление профилирования книжных серий, препринтов, обзоров.

*Профилирование организаций* распространяется на издательства, коллективных авторов, центры-депозитарии, места защиты диссертаций, так как для них есть основания предполагать специализацию по тематике. Так же, как для журналов, на основе статистической обработки выпущенных за последние три года информационных продуктов для некоторых организаций удалось выявить устойчивые тематические профили. Они занесены в описания объектов-организаций. По состоянию на

* Из расчета: 250 рабочих дней в году; 7 часов непрерывного труда в день.

начало 2005 г. из 22 668 объектов массива организаций тематические профили присвоены 280. Примеры в табл. 8.

*Таблица 7*

| Название издания | Страна | Вид | Тематический профиль |
|---|---|---|---|
| The Industrial Robot | Великобритания | журнал | Машиностроение |
| Mizunamishi kaseki hakubutsukan kenkyu hokoku = Bulletin of Mizunami Fossil Museum | Япония | журнал | Геология и горное дело |
| Итоги научно-исследовательской работы Государственного университета по землеустройству | Россия | книжная серия | География и геофизика<br>Экономика промышленности<br>Охрана окружающей среды |

*Таблица 8*

| Название организации | Страна | Тематический профиль |
|---|---|---|
| Издательский дом "Финансы и кредит" | Россия | Экономика промышленности |
| Издательство "Медиа Медика" | Россия | Биология |
| Opća bolnica Zadar | Венгрия | Биология |
| Výzkumný ústav pro práškovou metalurgii | Чехия | Металлургия |

*Профилирование изданий книжного типа* — задача иного порядка, поскольку до получения такого издания о нем ничего неизвестно (кроме выпусков книжных серий). Однако по результатам библиографической обработки отдельных видов изданий книжного типа можно сделать выводы о тематической направленности.

Для некоторых изданий библиографические описания содержат обязательные специфические элементы, прямо отражающие тематику. Так, авторефераты диссертаций имеют номер специальности, нормативные документы (стандарты, технические условия и др.) имеют классификационные коды. Поддержка в системе нормативно-справочной информации соответствующих классификаторов, оснащенных тематическими кодами ВИНИТИ, позволяет эффективно использовать эти сведения в целях автоматизации разметки*. Для оснащения классификаторов кодами разметки (причем не только 1-го уровня, но и более точными индексами) также привлекаются результаты статистической обработки информационных продуктов.

Депонированные рукописи по принятым правилам индексируется рубриками ГРНТИ, которые напрямую могут быть использованы для их тематической разметки.

Картографические издания естественно направлять в научную редакцию "География и геофизика".

*Профилирование научных мероприятий* производится непосредственно при порождении этих объектов: описание мероприятия индексируется рубриками ГРНТИ. Эти сведения напрямую могут быть использованы для разметки публикаций-докладов, входящих в раздел выпуска издания, соответствующий мероприятию.

Таким образом, *накапливаемые в массивах сериальных изданий и организаций тематические профили этих объектов, значения специфических элементов, явные индексы ГРНТИ используются для формирования результирующего штампа разметки 1-го уровня.* Алгоритм формирования учитывает возможность наличия тематического признака одновременно в нескольких элементах: выбор осуществляется в соответствии с иерархией (в порядке уменьшения значимости):

(а) индексы рубрик ГРНТИ;

(б) картографические издания;

(в) классификационные индексы (номер специальности авторефератов; индекс классификации нормативных документов);

(г) тематический профиль сериального издания (в т. ч. книжные серии);

(д) тематический профиль коллективного автора, места выполнения работы, места защиты диссертации;

(е) тематический профиль издательства.

**Использование аналитического описания.** Как отмечалось выше, авторская аннотация и ключевые слова могут служить для определения тематики. До недавних пор эти данные нельзя было использовать для автоматической разметки, так как они отсутствовали в компьютерной базе к нужному моменту. С возрастанием потока электронных публикаций и доли журналов, проходящих участок сканирования, разметчики могут получить доступ к аннотациям и ключевым словам, загруженным вместе с другими элементами библиографических описаний статей.

Известно много работ по автоматическому индексированию на основе полных текстов или фрагментов. Большинство из них опирается на развитые системы словарей.

В рамках АСКР проводятся подготовительные работы по формированию словарей ключевых слов и словосочетаний и оснащению их тематическими кодами. На первом этапе составляются русскоязычные словари. Источником данных служит многократно упомянутая статистическая обработка информационных продуктов ВИНИТИ, в результате которой создан массив слов и словосочетаний и получены частотные характеристики их употребления в парах с конкретными рубриками. Исходный массив включает около 300 тыс. различных слов и словосочетаний.

Начальная обработка массива состоит в очистке от мусора — выявлении одинаковых слов и словосочетаний, записанных по-разному из-за ошибок набора или несоблюдения единого порядка слов. Алгоритмы основаны на морфологическом анализе, дополненном вычислением близости слов и словосочетаний по хэш-ключам (по методу, упоминавшемуся в подразделе 2.1). В результате исходный массив разбивается на непересекающиеся подмножества — кластеры эквивалентных слов и словосочетаний, в каждом из которых определен главный

*То же надо сказать и о патентных документах, описания которых в качестве обязательного элемента содержат коды МПК. При включении патентов в рабочий поток АСКР этот факт необходимо учитывать и сразу реализовывать механизмы автоматической разметки.

термин. Таких кластеров оказалось более 150 тыс. Если исключить малоупотребляемые, то останется около 90 тыс.

Дальнейшие работы невозможны без привлечения специалистов научных редакций и общей методической поддержки. Используя списки рубрик, сопоставленные словам и словосочетаниям, каждому специалисту можно выделить для обработки список терминов, соответствующий его профилю. Деление общего массива на отдельные списки может быть осуществлено с различной точностью: грубо — по отделам, точнее — по выпускам Реферативного журнала, по конкретным рубрикам.

Содержанием работы специалиста должны быть действия по отсеву неинформативных терминов, объединению одинаковых по смыслу терминов, пополнению списка новыми терминами, установлению взаимоотношений между терминами и уточнению их тематических индексов. В настоящее время разрабатываются программные средства для обеспечения соответствующего инструментария.

В результате можно ожидать получение выверенных словарей слов и словосочетаний, снабженных тематическими индексами. Аспекты их приложения — не только автоматизация индексирования, но и построение поискового тезауруса.

## ЗАКЛЮЧЕНИЕ

Проект по созданию АСКР инициирован в 1997 г. В результате его выполнения в ВИНИТИ построена и успешно функционирует оригинальная система, реализующая автоматизированную технологию комплектования и обработки входного потока литературы, обеспечивая взаимодействие десятков пользователей на всех стадиях технологического процесса. Имея в своем арсенале средства управления продвижением физических экземпляров по участкам и адекватного отражения технологического процесса в базе данных, АСКР способна перерабатывать большой поток литературы, поступающей в информационный центр. Опытно-промышленная эксплуатация системы началась в 1999 г.; с 2000 г. она эксплуатируется в промышленном режиме.

АСКР основана на реляционной модели и поддерживается средствами СУБД Microsoft SQL-Server. Аккумулируя содержание статьи, перечислим наиболее важные **решения, положенные в основу архитектуры АСКР**.

В части построения базы данных:

декомпозиция описания публикации, выразившаяся в выделении самостоятельных *объектов* (издания, экземпляры, выпуски, разделы, организации, персоны, научные мероприятия), которые взаимодействуют в результирующем описании обрабатываемого документа;

определение *классов* однотипных объектов — с описанием их общих свойств, правил поведения и связей с объектами других классов;

введение над стандартными средствами SQL понятия *массива данных* для хранения описаний объектов одного класса в сочетании с нормализацией отношений на уровне объектов.

В части построения программного комплекса:

централизованная полнофункциональная *поддержка алфавита* — с обеспечением контроля и визуализации специальных символов при выполнении любых операций над текстами;

общесистемная *проверка структуры описаний* объектов и корректности ссылок при организации взаимосвязей;

специальные средства для *выявления дублей* объектов в массивах данных;

реализация технологического процесса в моделях *жизненных циклов* объектов обработки с гарантией целостности технологических цепочек от порождения объектов до их утилизации;

построение *автоматизированных рабочих мест* путем составления комбинаций из множества относительно независимых клиентских приложений.

В части обеспечения согласованной и управляемой работы всех компонентов АСКР исключительно важна *система нормативно-справочной информации*, оснащенная соответствующим методическим и программным обеспечением.

Таким образом, АСКР представляет собой некое подобие объектно-ориентированной системы, реализованной средствами реляционной СУБД.

Наверное, некоторые из перечисленных решений по отдельности не составляют открытий для опытных разработчиков. Тем не менее, кажется уместным сгруппировать их в заключении, чтобы на примере АСКР подчеркнуть тот факт, что последовательное воплощение указанных моделей данных и методов манипулирования данными является не только **необходимым условием работоспособности большой системы, но и открывает возможности для ее развития**.

В качестве примеров наращивания функциональности АСКР были рассмотрены проекты: "Обработка публикаций из электронных источников", "Научные мероприятия", "Автоматизация тематической разметки". В 2002–2003 гг. эти разработки вышли из стадии проектирования и эксплуатируются в опытно-промышленном режиме.

На очереди стоят задачи оптимизации и развития технологии: ресурсосберегающая обработка депонированных рукописей; автоматизированная обработка патентов (с использованием машиночитаемых описаний, которые можно получить от отечественных и зарубежных центров обработки патентных документов); расширение фронта аналитической обработки; углубление средств автоматической разметки на уровень документов; нормализация представления подсерий в массиве сериальных изданий; организация параллельной обработки экземпляров-дублей по разным технологическим маршрутам.

Возможность успешной реализации таких проектов в большой степени обусловлена архитектурой АСКР, которая позволяет органично встраивать новые объекты и процессы в существующую технологию.

Наконец, хочется отметить, что АСКР способна решать задачи, выходящие за рамки основной цели своего функционирования (см. [1], Введение). Т. е., не только снабжать научные редакции документами для содержательной обработки и последующего помещения в Реферативный журнал, но и **самостоятельно производить некоторые информационные продукты**, которые могут быть получены в результате первичной обработки литературы и представлять интерес для потребителей.

В первую очередь, речь идет о Каталоге поступлений в совокупности с Массивом документов.