ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК [002:004]:001.83 ВИНИТИ

В. С. Егоров, А. В. Пожидаев, Т. Н. Чернобровская

Систематизация и использование сведений о научных мероприятиях в автоматизированной технологии ВИНИТИ

Рассматриваются автоматизированная система сбора, обработки, систематизации и представления в выходных информационных продуктах ВИНИТИ описаний научных мероприятий, а также наиболее важные решения: формализация, структура и технологическая схема информационного массива описаний научных мероприятий.

Научные коммуникации являются основой и транспортной средой науки. Без общения, без обмена опытом и мнениями невозможно достижение научных результатов. Если до конца 20-го столетия одной из наиболее эффективных форм научного общения были конференции, симпозиумы и другие научные мероприятия с прямым общением участников, то сейчас наибольшее значение приобретают научные коммуникации с использованием технологий Интернета (электронная почта, электронные журналы, электронные конференции и т. д.). Однако эти технологии не могут полностью заменить непосредственное общение ученых и специалистов, поэтому значение традиционных научных мероприятий остается востребованным.

Для того чтобы быть на передовом крае науки, специалист должен владеть информацией о проводимых мероприятиях по профилю его научной деятельности. Учитывая то, что по очень грубым оценкам сейчас в мире ежегодно проводится свыше 5 тысяч научных мероприятий (фундаментальные и технические науки), выбор и самостоятельное формирование подобного пользовательского списка представляется достаточно трудоемким процессом. До появления Интернета единственным источником сводных, систематизированных сведений о научных мероприятиях были печатные издания, наиболее ценными среди них — сводные обобщающие материалы. Практически все информационные центры издают или издавали подобные информационные материалы, научные общества и организации рассылали списки планируемых мероприятий. В ВИНИТИ таким изданием является издающийся с 1960 г. "Бюллетень международных научных съездов, конференций, конгрессов, выставок". Сейчас автоматизированные и телекоммуникационные технологии активно вытесняют печатные издания, которые становятся производным элементом от самостоятельного развивающихся электронных технологий, основанных на поддержании банков данных. Указанные банки данных становятся основным информационным продуктом, при этом пользователь получает принципиально новые возможности, он может активно работать со всем ретроспективным объемом данных, иметь доступ к оперативно подготавливаемой информации.

Проведенный в течение 2005 г. анализ показал, что вопросами сбора и систематизации данных о проводимых научных мероприятиях занимается достаточно большое количество организаций, большинство из них предоставляет через Интернет бесплатный доступ к созданным ресурсам. Примером таких агрегаторов в России являются Министерство образования и науки Российской Федерации (Федеральное Агентство по Науке и Инновациям) http://www.fasi.gov.ru и Информационная система "Наука и Инновации" http://www.rsci.ru. Среди зарубежных источников можно упомянуть сайты http://www.allconferences.com, http://www.eventseye.com/, http://atlas-conferences.com/.

Многие поисковые машины имеют разделы, связанные с информацией по конференциям. Например, самый крупный поисковик Google имеет специализированный раздел http://www.google.com/Top/Science/Conferences/.

Однако более детальный анализ web-сайтов (порталов) со списками планируемых мероприятий обнаружил, что практически каждый из них охватывает сравнительно небольшой объем информации (в большинстве случайной или очень узкой тематики), информация плохо структурирована, пользовательский интерфейс ориентирован на англоязычного пользователя. Поэтому в ВИНИТИ началась работа по созданию массива данных по научным мероприятиям, информация о которых находится в сфере внимания Института — ведущего информационного центра страны в области научно-технической информации. Не менее важным фактором необходимости создания такого продукта является формирование механизма для обеспечения наиболее полных и ценных поступлений

во входной поток Института материалов прошедших конференций — объемной и важной его составляющей. Следует отметить, что в большинстве случаев материалы конференций — это бесплатный информационный продукт, что часто становится решающим фактором при комплектовании входного потока научно-технической информации (НТЛ). Наличие оперативно поддерживаемого массива планируемых научных мероприятий позволит автоматизировать процессы взаимодействия с оргкомитетами для получения соответствующих материалов.

Настоящая публикация излагает основные результаты работ, выполненных Центром разработок информационных систем (ЦРИС) ВИНИТИ в 2004–2005 гг. За этот период осуществлена работа по созданию автоматизированной системы сбора, обработки и систематизации данных для формирования массива научных мероприятий; начата опытная эксплуатация её пускового варианта (разработано программное и технологическое обеспечение); осуществлен перевод на новую технологию подготовки "Бюллетеня международных научных съездов, конференций, конгрессов, выставок", включая оригинал-макет. Создаваемый массив научных мероприятий будет иметь очень важный поисковый аспект, обеспечивая взаимосвязь между планируемыми конференциями и отслеживание выходящих публикаций и других материалов по результатам работы конференций. Для облегчения поиска мероприятия будут систематизированы по тематике в соответствии с Государственным рубрикатором НТИ (ГРНТИ).

ФОРМАЛИЗАЦИЯ ОПИСАНИЯ НАУЧНОГО МЕРОПРИЯТИЯ

Основным информационным объектом создаваемой системы сбора, обработки и систематизации данных для формирования массива научных мероприятий является их описание. Поэтому следует более точно определить, что понимается под представлением научного мероприятия. Следуя общепринятой терминологии, будем считать, что: *Научное мероприятие — это научная конференция, симпозиум, семинар, выставка и т. п., проводившееся или планирующееся к проведению.*

При решении задач формализации были поставлены и решены следующие основные задачи:
- унификация описаний мероприятий для организации удобных поисковых механизмов;
- создание механизмов для обнаружения дублирования при массовом вводе информации;
- создание корректного механизма установления взаимоотношений между различными мероприятиями.

В процессе решения этих задач учитывалась необходимость обеспечения интеграции разрабатываемой системы в существующие технологические процессы.

Создаваемый массив должен обеспечить хранение унифицированных сведений о научных мероприятиях, попадающих в сферу интересов ВИНИТИ, — от получения материалов до выпуска периодического издания и рассылки сообщений с анонсами мероприятий.

В исходном описании любого научного мероприятия чаще всего содержатся лишь три обязательных элемента — оригинальное название, место и дата проведения. Однако для обеспечения эффективного поиска и систематизации такой ограниченный набор элементов недостаточен, поэтому было принято решение дополнить перечень элементов описания научного мероприятия как элементами, несущими дополнительную информацию о мероприятии, так и элементами-классификаторами, определенными в процессе анализа существующих методик описания мероприятий в печатных и электронных изданиях и необходимыми для реализации задач поиска и систематизации. Расширенный перечень элементов описания выглядит следующим образом:
- наименование мероприятия — оригинальное;
- порядковый номер к наименованию;
- наименование мероприятия на русском языке (в том случае, если язык библиографического описания не русский);
- наименование мероприятия параллельное (для редких языков);
- дата проведения мероприятия: начало — год, месяц, день и окончание — год, месяц, день;
- страна проведения;
- место проведения мероприятия (населенный пункт);
- организатор(ы) мероприятия;
- тип мероприятия;
- форма проведения мероприятия;
- язык библиографического описания;
- язык параллельного библиографического описания;
- характер мероприятия;
- географический охват мероприятия;
- состав участников;
- адрес проведения мероприятия (не населенный пункт и (или) конкретное место в населенном пункте);
- рабочий язык;
- ключевые слова;
- краткое резюме;
- материалы мероприятия;
- тематики мероприятия;
- секции мероприятия;
- связи с описаниями других мероприятий;
- актуальность описания.

Таким образом, описание мероприятия содержит 24 элемента. Безусловно, заполнение всех элементов не обязательно и определяется в каждом конкретном случае полнотой и доступностью в источнике сведений о мероприятии.

Основополагающим разделением мероприятий был выбран признак серийности — мероприятие может быть представлено в массиве либо *разовым*, либо *серийным*. Разовое мероприятие — достаточно редкий случай. Это мероприятие, которое проводилось однократно и не планируется к проведению в дальнейшем. Например — выставка, посвященная 1000-летию Казани.

Серийное мероприятие — более распространенный случай. Серийными считаются мероприятия, содержащие хотя бы один из следующих признаков:
- порядковый номер в названии мероприятия;
- год проведения в названии мероприятия;
- специальные термины в названии мероприятия — "сессия", "чтения", "годичная", "ежегодная", "постоянно действующая" и т. п. Данные термины объединены в специализированный словарь, который в перспективе будет использован

при автоматическом определении серийности мероприятия.

Помимо разделения "разовое"/"серийное", в процессе обработки сведений о мероприятиях возникла необходимость в еще одном типе мероприятия — *обобщенное мероприятие*. Обобщенное мероприятие — это описание мероприятия, никогда не проводившегося в реальности, но необходимое для определения последовательности мероприятий. Например, в 2003 г. проводилось мероприятие — SAS European Users Group International (SEUGI-21), а в 2004 г. — SAS Forum Internationa — 2004. Визуально, мероприятия не связаны, однако на сайте SAS Forum есть упоминание "<...> прозвучало на международном форуме SAS — SAS Forum International 2004 <...>, ранее известной как SeUGI (SAS European Users Group International), [конференция изменила в этом году свое название]". Ввод обобщенного описания "SAS Forum" и связывание с ним перечисленных выше мероприятий, позволит найти все предыдущие мероприятия, проинформировать пользователя, проводящего поиск по старому названию, о смене названия, а также позволит найти *лакуны* — мероприятия, информация о которых отсутствует в массиве мероприятий (т. е. если в массиве есть описания серии с 1-й по 6-ю и 8-й конференций, то необходимо найти информацию о 7-й).

В дальнейшем было решено разделить понятие "обобщенное мероприятие" на:

обобщенное монотематическое описание — собственно описание обобщенного мероприятия, связанного с серийными мероприятиями;

обобщенное политематическое описание — обобщенное описание, связанное с несколькими обобщенными монотематическими описаниями.

Необходимость подобного деления на данный момент отсутствует, однако подобная иерархическая структура создает задел на будущее, а также демонстрирует гибкость разработанной структуры при применении ее к любым описаниям мероприятий. Проиллюстрируем возможный случай, где будет необходимо многоуровневое деление обобщенных описаний, на рис. 1.

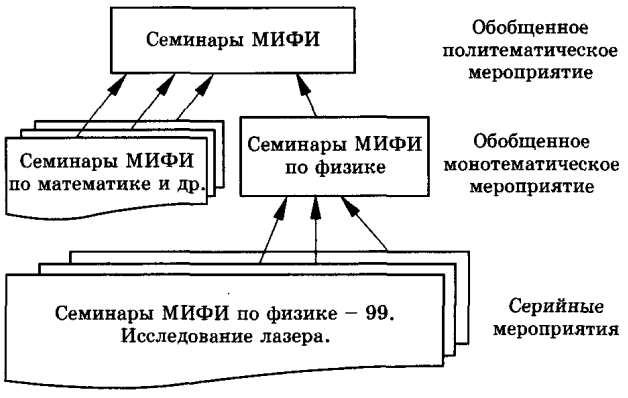

Рис. 1. Иерархия обобщенных описаний

СТРУКТУРА ИНФОРМАЦИОННОГО МАССИВА

Массив научных мероприятий реализован в виде таблиц реляционной СУБД MS SQL Server, среди которых основная таблица EVENTS_UP, 15 вспомогательных таблиц и 26 словарей. Основная таблица EVENTS_UP хранит большую часть сведений о мероприятии и выполняет агрегирующую функцию. Помимо нее, для хранения всех 24 элементов описания мероприятия используются еще 5 таблиц (рис. 2).

Необходимо отметить, что дата проведения мероприятия хранится в основной таблице в отдельных полях, что вызвано необходимостью хранить (и уточнять) сведения о мероприятиях, в качестве даты проведения которых в источнике указаны лишь месяц и год (либо только год).

Основными словарями с классификаторами в массиве научных мероприятий являются:

- тип мероприятия ("Описание разового мероприятия", "Описание обобщенного мероприятия", "Описание серийного мероприятия");
- форма проведения ("Реальная конференция", "Виртуальная конференция", "Заочная конференция");
- характер мероприятия ("Выставка", "Научное мероприятие");
- вид географического охвата ("Международная", "Национальная", "Региональная", "Межрегиональная", "Городская", "Областная");
- вид состава участников ("Научный состав", "Молодые ученые", "Специалисты-практики", "Учащиеся");
- вид носителя материала ("Книга", "Журнал", "веб-сайт Интернета", "Информационный лист", "Электронное издание").

Обработка сведений о мероприятиях по этим классификаторам осуществляется в основной программе взаимодействия с массивом мероприятий "PdpCon", как в ручном, так и в автоматизированном режиме, для чего словари содержат переводы терминов на основные языки.

Основным инструментом формирования и поддержания массива научных мероприятий является программа PdpCon. Ее задачи: создание, редактирование и удаление описаний мероприятий. Однако, помимо этого, программа позволяет выполнять:

- создание описаний мероприятий в автоматизированном режиме;
- многокритериальный поиск в массиве мероприятий;
- создание оригинал-макета "Бюллетеня международных научных съездов, конференций, конгрессов, выставок";
- создание информационных листков для различных информационных продуктов ВИНИТИ;
- автоматизированную обработку сведений о научных мероприятиях из регистрационного массива опубликованных материалов (ретро-фонд);
- выгрузку и загрузку информации для программы перевода;
- сбор статистической информации в массиве мероприятий.

С целью облегчения работы по вводу в систему информации о мероприятиях в Центре разработок информационных систем (ЦРИС) ВИНИТИ было разработано вспомогательное программное обеспечение:

- программа включения в массив мероприятий результатов перевода названий иностранных мероприятий;
- программа маркирования анонсов, размещенных в Интернете;
- программа упрощенного ввода описаний мероприятий;

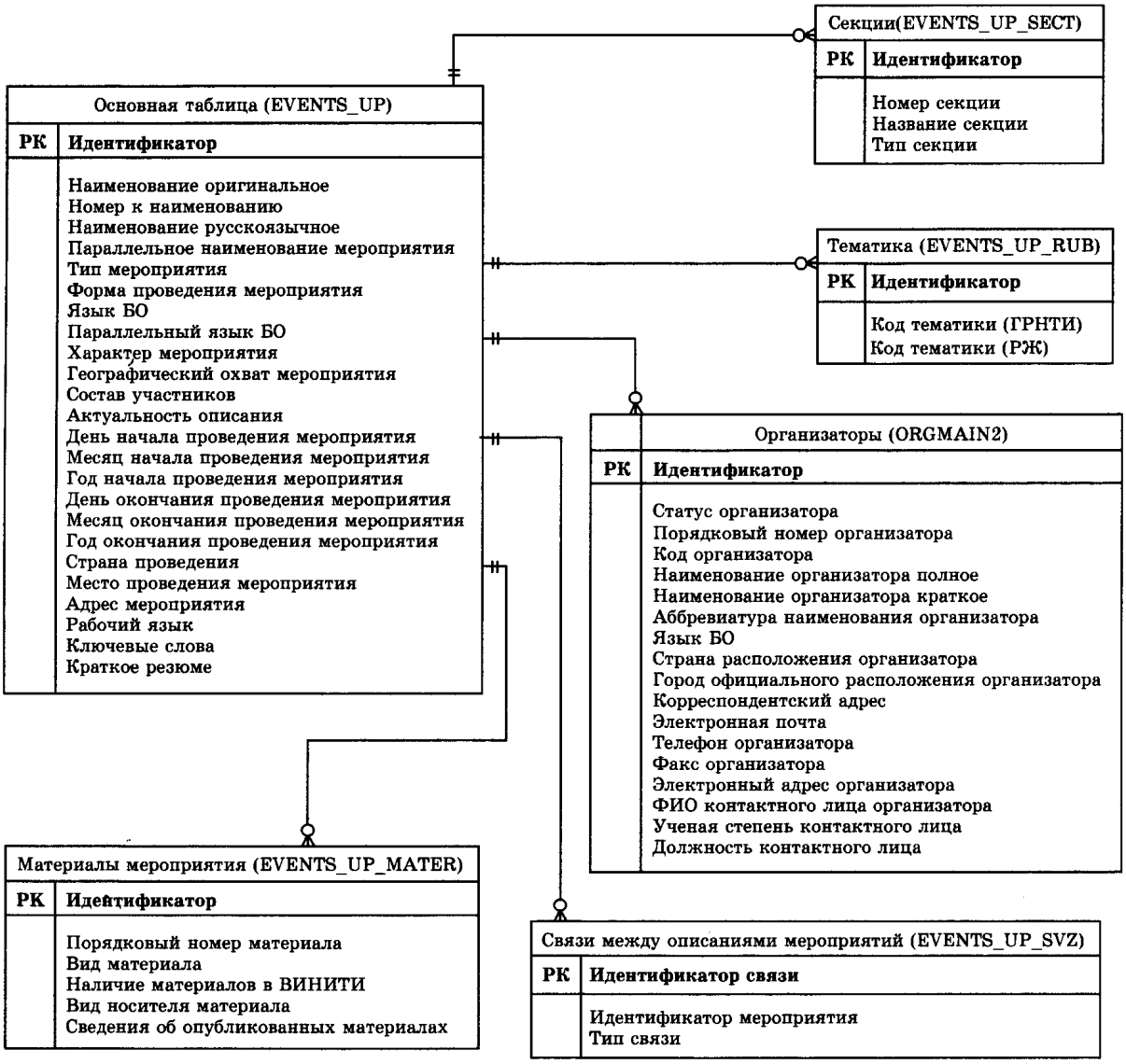

Рис. 2. Основные таблицы массива научных мероприятий

• программа, выполняющая основные статистические запросы к массиву мероприятий.

В основной программе "PdpCon" реализована полноценная работа со связями между описаниями мероприятий — начиная от их создания в ручном или автоматизированном режиме и заканчивая учетом этих связей при формировании выходных продуктов (например, несколько мероприятий, проводившихся совместно, хранятся в массиве мероприятий по отдельности, но в "Бюллетене международных научных съездов, конференций, конгрессов, выставок" они будут отображены одной записью).

Механизм связей позволяет, помимо реализации иерархии описаний (см. выше), реализовывать связи между мероприятиями. Причем связи могут быть как известные изначально (мероприятия, проводимые совместно, или одно в рамках другого), так и связи, установленные при обработке (связь "актуальное описание — устаревшее описание"). Это позволяет одновременно хранить в массиве и отдельные серии мероприятий, и связи между мероприятиями из разных серий.

Для хранения связей был разработан следующий механизм (см. рис. 3).

Были созданы две таблицы — таблица с перечнем связей (askeve.EVENTS_UP_SVZ_1) и таблица связей между описаниями (askeve.EVENTS_UP_SVZ). Необходимость двух таблиц обусловлена тем, что наиболее рациональная — иерархическая — структура невозможна без выделения "главной" записи, а задача выделения главной записи среди нескольких совместно проводившихся мероприятий значительно снизит скорость работы оператора, вводящего данные о мероприятиях. При такой структуре могут возникнуть дополнительные вопросы к оператору: при удалении "главной" — ему необходимо будет среди оставшихся записей выделить новую "главную" запись.

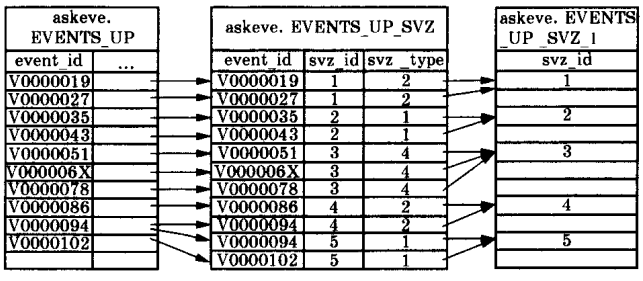

Рис. 3. Возможные связи между описаниями научных мероприятий в массиве мероприятий

Рассмотрим пять наиболее вероятных случаев создания связей:

1. Создание связи между мероприятиями, проводившимися совместно (V0000019 и V0000027) — в этом случае создается связь "1" с соответствующим комментарием. После чего к этой связи подключаются оба мероприятия с указанием типа связи "2" ("совместное").

2. Создание связи между устаревшим и новым описанием мероприятия (V0000035 и V0000043) — в этом случае создается связь "2" с соответствующим комментарием. После чего к этой связи подключаются оба мероприятия с указанием типа связи "1" ("новое — устаревшее"). Устаревшее описание в данной ситуации определяется по признаку "устаревшее" в описании мероприятия.

3. Создание обобщенного описания (V0000051 на основании нескольких серийных — V000006X и V0000078) — в этом случае создается связь "3" с соответствующим комментарием. После чего к этой связи подключаются все мероприятия с указанием типа связи "4" ("обобщенное — серийное"). Обобщенное описание в данной ситуации определяется по типу мероприятия.

4. Создание связи между устаревшим и новым описанием мероприятия (V0000094 и V0000102), при этом уже существует связь между мероприятиями, проводившимися совместно (V0000086 и V0000094) — в этом случае создается связь "2" с соответствующим комментарием. После чего к этой связи подключаются оба мероприятия с указанием типа связи "1" ("новое—устаревшее"). Устаревшее описание в данной ситуации определяется по признаку "устаревшее" в описании мероприятия. Проводится поиск связей "устаревшего" описания и его связи переносятся на новое описание.

5. Удаление одной из записей, состоящих в связи, происходит через удаление соответствующей записи в таблице связей (EVENTS_UP_SVZ). При этом происходит проверка: если запись в таблице с перечнем связей (EVENTS_UP_SVZ_1) больше не имеет ссылок на таблицу связей (EVENTS_UP_SVZ), то удаляется и сама связь (т. е. запись в askeve.EVENTS_UP_SVZ_1). При удалении, например, обобщенного описания, связанные с ним серийные описания сохранят связь между собой и могут быть в любой момент присоединены к новому обобщенному описанию.

ТЕХНОЛОГИЯ НАПОЛНЕНИЯ ИНФОРМАЦИОННОГО МАССИВА

Наполнение массива мероприятий осуществляется по технологической схеме, представленной на рис. 4.

Информацию, поступающую из различных подразделений ВИНИТИ для пополнения массива мероприятий, можно разделить на три вида:

1. Анонсы мероприятий на бумажных носителях (справочники, поставляемые информационными центрами, "Перечень международных конференций РАН" и т. п.); зарубежные периодические издания (Meetings on Atomic Energy); информационные листы с анонсами; информация, поступающая из подразделений ВИНИТИ.

Анонсы на бумажных носителях являются традиционным источником данных для формирования информационного продукта ВИНИТИ — "Бюллетеня международных научных съездов, конференций, конгрессов, выставок" и технология их получения и обработки достаточно отработана.

2. Анонсы мероприятий на электронных носителях (электронные библиотеки; информационные ресурсы; веб-страницы организаторов мероприятий и т. п.). Это новый источник информации и разработка технологии его использования происходит в настоящий момент.

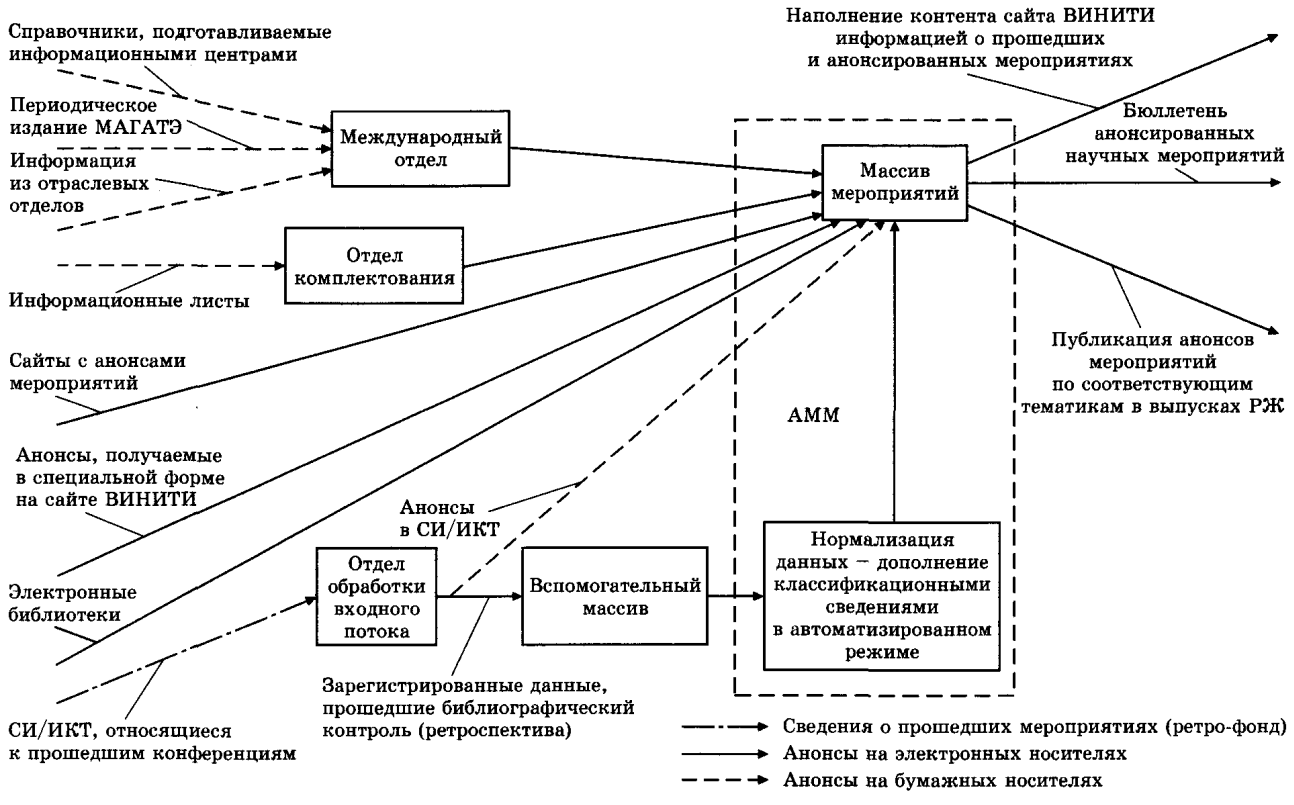

Рис. 4. Наполнение массива мероприятий и дальнейшее использование этой информации

3. Сведения о прошедших мероприятиях (информация о прошедших мероприятиях, выделенная в процессе обработки сериальных изданий и изданий книжного типа (СИ\ИКТ), на основании которых и были выпущены данные издания). Этот источник используется для отражения полного цикла жизни мероприятия (от получения анонса до публикации его результатов) и поддержания полноты цепочек серийных мероприятий (заполнение "лакун" — см. выше).

Дальнейшей обработкой информации из регистрационного массива опубликованных материалов (ретро-фонд), выделением анонсов из СИ\ИКТ (эта работа не входит в задачи отдела обработки входного потока), получением анонсов с сайтов в сети Интернет (в том числе, с формы регистрации на сайте ВИНИТИ) и из электронных библиотек занимаются — *администраторы массива мероприятий* (АММ). Основная функция АММ — поддержка и обслуживание массива мероприятий, однако помимо этого, АММ играет интегрирующую роль в процессе сбора информации о мероприятиях из разных отделов ВИНИТИ, а также АММ занимается обработкой сведений о тех мероприятиях, которые не проходят обработку в отраслевых отделах.

Для облегчения работы по вводу информации в систему было разработано специальное программное обеспечение. В частности, для отдела комплектования было разработано программное обеспечение и новый коммуникативный формат, позволившие реализовать технологическую цепочку "выгрузка сведений о мероприятиях на иностранных языках, не имеющих русского перевода — формирование файла с информацией — перевод, осуществляемый на любых компьютерах (в том числе, не подключенных к серверам ВИНИТИ) — автоматическая загрузка", а для администраторов массива разработано программное обеспечение, позволяющее реализовать технологию *маркирования* — выделения информации в анонсах мероприятий, размещенных в Интернете, цвето-стилистическими схемами, что позволяет осуществлять автоматизированный ввод маркированной информации в массив мероприятий.

Основным информационным продуктом, выпускаемым в настоящее время с использованием массива научных мероприятий, является обновленный "Бюллетень международных научных съездов, конференций, конгрессов, выставок", который отличается от своего предшественника тем, что имеет перевод на русский язык иностранных мероприятий, указатель тематик по кодам ГРНТИ, указатель организаторов и мест проведения мероприятий, а также большим охватом.

Освоен выпуск информационных листков и для других выходных продуктов ВИНИТИ, в этом случае анонсы отбираются по конкретным тематическим разделам.

В дальнейшем массив научных мероприятий может быть доступен интернет-пользователям, подписчикам "Бюллетеня международных научных съездов, конференций, конгрессов, выставок", подписчикам различных информационных продуктов ВИНИТИ, отделу комплектования Института (для наполнения входного потока).

ТЕКУЩЕЕ НАПОЛНЕНИЕ МАССИВА

На конец 2005 г. в массиве содержались сведения о 2253 мероприятиях — из них: будущих — 587 мероприятий, состоявшихся — 1666 мероприятий.

Объем входного потока составляет примерно 600 мероприятий в месяц, которые в процентном соотношении делятся следующим образом:

- сведения о прошедших мероприятиях из отдела обработки входного потока — 50%;
- анонсы в Интернете, обрабатываемые АММ, — 20%;
- анонсы, поступающие из международного отдела, — 20%;
- информационные листы, поступающие из отдела комплектования, — 10%.

Состояние массива мероприятий можно оценить, рассматривая различные его срезы, например (см. Рис. 5, 6, 7):

- по форме проведения — конференции (2170)/ выставки (83);
- по виду географического охвата — Международные (953)/ Российские (280)/ Региональные (89)/ Межрегиональные (23)/ Городские (7)/ Областные (5);
- по странам проведения — Россия (1170)/ США (222)/ Франция (90)/ Великобритания (79)/ Германия (65);
- по виду материалов, содержащих информацию о мероприятии — Доклад (898)/ Анонс (744)/ Тезисы (211);
- по тематикам ГРНТИ (одно мероприятие может включать несколько тематик) — Биология (402), Медицина и здравоохранение (337), Физика (334), Химия (242), Экономика. Экономические науки (224), Охрана окружающей среды. Экология человека (208), Автоматика и телемеханика. Вычислительная техника (182), Математика (121), Геология (119), Транспорт (96).

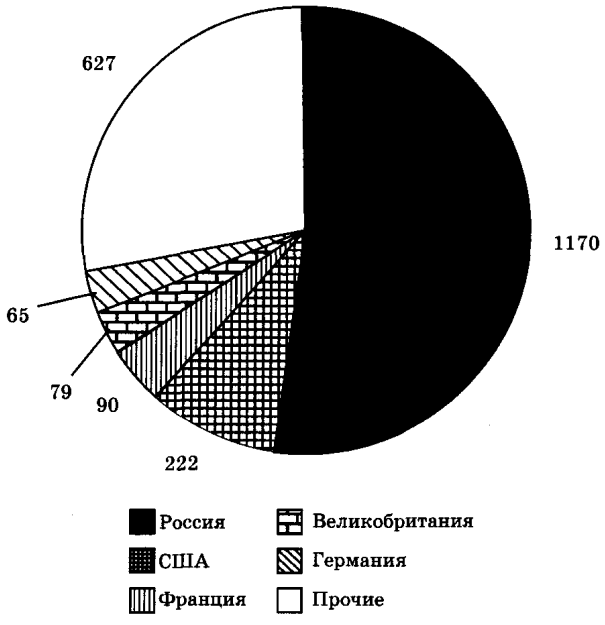

Рис. 5. Классификация имеющихся в массиве научных мероприятий по странам проведения