АНАЛИЗ ДАННЫХ ДЛЯ ГЕОИНФОРМАЦИОННЫХ СИСТЕМ

В. В. Попович, Д. В. Бербенев, А. В. Панькин, О. В. Смирнова

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН, 14 Линия ВО, д. 39,Санкт-Петербург, 199178

<popovich@mail.iias.spb.su>
<http://www.oogis.ru>

УДК 004.8

Попович В. В., Бербенев Д. В., Панькин А. В., Смирнова О. В. **Анализ данных для геоинформационных систем** // Труды СПИИРАН. Вып. 6. — СПб.: Наука, 2008.

Аннотация. *Геоинформационные системы (ГИС) различного назначения и масштаба являются прежде всего интерфейсом для доступа к данным и управлением их преобразования в интересах пользователя. Специфической чертой ГИС данных является то, что они имеют, как правило, географическую привязку, то есть связаны с географическими координатами.*

В данной статье рассматривается весь цикл преобразования данных для ГИС, которые используются в качестве интерфейса для систем поддержки принятия решений. Как показал опыт использования ГИС и анализ доступной литературы, класс данных систем чрезвычайно велик. Но особый интерес представляют системы поддержки принятия решений на базе ГИС в таких областях, как управление муниципальными образованиями, управление транспортом (морским, наземным), управление системами мониторинга различного назначения и масштаба.

В общем виде процесс доступа и преобразования данных рассматривается с точки зрения трех уровней абстрагирования: гармонизированные, интегрированные и сплавленные данные. Данные уровни имеют физическую основу и отражают, по сути, процесс прохождения информации от средств измерения вверх по иерархии. В статье рассматриваются только основные положения, без детализации внутренних процессов, присущих каждому уровню. — Библ. 6 назв.

UDC 004.8

Popovich V. V., Berbenev D. V., Pankin A. V., Smirnova O. V. ***Data for geoinformation systems*** // SPIIRAS Proceedings. Issue 6. — SPb.: Nauka, 2008.

Abstract. *Various purpose and scale geoinformation systems (GIS) are first and foremost interfaces to arrange for data access and user–friendly control of their transformation. As a rule GIS data specifics consists in their geographic reference (bind) i.e. reference to geographic coordinates.*

The here presented research proposes to consider the complete GIS data transformation cycle to further implement these GIS as interfaces supporting decision making systems. Judging from the so far accumulated experience of GIS implementation and based on the available literature study it could be said that the class of these systems is extremely large. However, a special interest is now drawn to the GIS based decision making systems for managing the municipal aggregations, transportations (maritime, ground); controlling various purposes and scales monitoring systems.

In general case the process of data access and transformation is assumed to be considered in the light of three abstracting levels: harmonized, integrated and fused data. The above mentioned levels possess a physical basis and really reflect the information flows propagating from measuring devices up the hierarchy. The paper is only focused on main statements and does not go into details specifying each level. — Bibl. 6 items.

1. Введение

Прежде всего приведем общеизвестное определение понятия «данные». Данные — это часть информации обычно формализованная в виде, пригодном для обработки и передачи. Данные могут быть представлены в виде чисел или текста на листе бумаги, битов или байтов, содержащихся в электронной памяти или сведений, хранящихся в уме человека. Метаданные — это информация о

данных. Метаданные описывают как, когда, где и кто собрал данные и как эти данные были отформатированы [1].

В данной статье мы не будем заострять внимание на отличии понятий «данные» и «информация», хотя во многих исследованиях и на обыденном уровне они воспринимаются как тождественные. Но все же мы будем говорить только о данных.

Значительный интерес к понятиям «информация» и «данные» возник с появлением Интернета, а еще ранее с появлением геоинформационных систем (ГИС). Возможно, что разработчики ГИС одними из первых столкнулись с проблемами использования разнотипной и достаточно объемной информации в реальном или в близком к реальному масштабе времени.

Данная проблема становится более острой в случае, когда ГИС используется в качестве интерфейса в системах поддержки и принятия решений в таких предметных областях, как управление муниципальными образованиями, управление движением судов и наземным транспортом, экологический мониторинг и многих других. Сегодня легче назвать те предметные области, где не используются ГИС, чем перечислить все те, где они интенсивно используются или планируются к использованию в ближайшее время.

Разнообразие ГИС привело к возникновению огромного числа информационных ресурсов для обеспечения ГИС необходимыми данными. Основу каждого ресурса, как правило, составляет определенная модель данных или так называемый формат представления данных. Но проблема заключается в том, что существующие форматы данных и основанные на них ресурсы не обеспечивают, за исключением специальных случаев, всех информационных потребностей современных систем поддержки и принятия решений (СППР). Таким образом, возникает задача группирования на концептуальном уровне возможных источников данных.

В данной статье предлагается выделить три группы или три типа данных: гармонизированные, интегрированные и сплавленные (слитые) данные. Такая группировка имеет смысл для понимания следующих аспектов процессов использования и преобразования данных. Это

- определение типа данных (измеренные данные, предварительно обработанные данные, экстраполированные и/или интерполированные данные и т. д.);
- определение источника данных и параллельно качества данных и степени доверия к ним;
- возможность использования данных для решения конкретных задач;
- возможность дальнейшего преобразования данных (изоморфные преобразования интегрированных данных, как правило, невозможны).

Перечисленный выше список является далеко не исчерпывающим, но позволяет объяснить идею или цель концептуальной декомпозиции данных.

2. Гармонизация информации

Данный процесс предполагает определение основных понятий и их взаимоотношений (онтологии) по соответствующим предметным областям и/или сферам ответственности. Например, разделение может быть выполнено по существующим областям знаний: гидроакустика, гидрометеорология, радиолокация, теория поиска и т.д. Гармонизация информации решает следующие основные задачи:

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

- обеспечение доступа к возможно большему числу первичных источников информации;
- возможность преобразования информации в удобный для пользователя вид (декодирование, распознавание, перевод и т.д.);
- обеспечение доступа к существующим информационным ресурсам.

Гармонизация в широком смысле может трактоваться как стандартизация данных.

Обеспечение доступа к первичным источникам информации может быть решено на двух уровнях — аппаратном и программном.

В интересах ГИС в общем виде можно предположить следующие типы источников информации. Это

- неформализованная информация (обычный текст, растровая графика, фото и т.д.);
- формализованная информация (например, в XML формате);
- результаты измерений в формализованном виде (в текстовом и цифровом);
- базы данных различных форматов;
- картографическая информация в специализированных форматах;
- информация о среде в различных специализированных форматах.

Графически гармонизация информации может быть иллюстрирована на рис. 1.

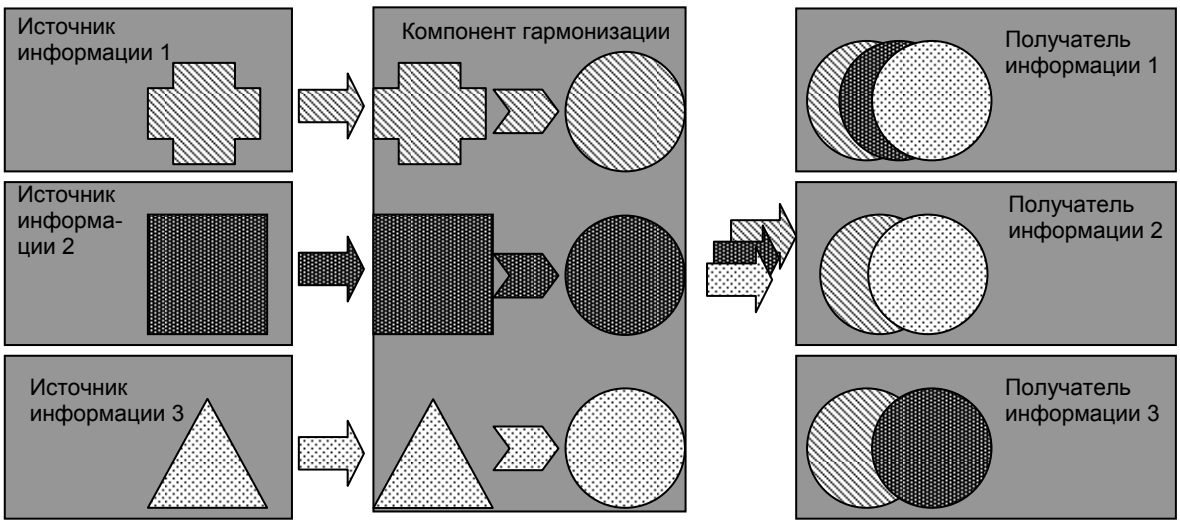

Рис. 1. Гармонизация информации.

Как видно из рис. 1, доступ к каждому источнику информации осуществляется, как правило, по разным протоколам, способам и/или механизмам. В качестве примера можно привести доступ к Интернет–ресурсам, базам данных, GPS, GSM данным, книги, статьи и т.д. Суть гармонизации заключается в реализации понятных принципов и механизмов доступа к информации, их унификации и сужения количества типов. Для примера можно привести коды Всемирной метеорологической организации (ВМО). В настоящее время основная часть данных рассылается в виде телеграмм, построенных на принципах передачи данных факсимильными телеграммами что очень сильно затрудняет их обработку и дальнейшее использование. В настоящее время ВМО проводит работу по приведению данных к единому XML–формату, что значительно упростит работу с такими данными.

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Отличительной чертой процесса гармонизации информации является то, что результат гармонизации ориентирован на большое число потребителей.

Как отмечается в [2], гармонизация информации о внешней среде имеет существенное значение как в региональном, национальном, европейском и глобальном контексте, так как ее определяют в первую очередь:

- глобальный мониторинг поверхности Земли, природных ресурсов и других данных для обработки и реализации в соответствии с Киотским протоколом;
- политика по окружающей среде в Европе, включая защиту внешней среды, развитие городов, защиту от природных катаклизмов;
- риск от вредных выбросов, геофизических опасностей и технологических рисков;
- международное сотрудничество, политика безопасности посредством разработки карт и систем поддержки принятия решений.

Разработка региональной, национальной, европейской или глобальной инфраструктуры пространственных данных ставит требование доступности информации и взаимообмена. Отсюда вытекают требования стандартизации и разработки соответствующих технологий. Реальное состояние дел ставит вопрос о доступности информации между различными сообществами, тем самым стимулирует развитие усилий по гармонизации данных через разработку общей геомодели данных.

Разработка общей геомодели данных позволит пользователям обращаться к различным источникам данных и использовать различное программное обеспечение в собственных интересах.

Геоинформационное сообщество Европы поставило задачу создания открытой организации по координации усилий по гармонизации информации.

По инициативе British Geological Survey (BGS) and of the Geological Survey of Canada (GSC) состоялась встреча в Единбурге, в ноябре 2003 г. На данной встрече присутствовали представители 15 геологических агентств из различных стран и континентов (Европа, Америка, Азия, Австралия). На данной встрече была создана рабочая группа по разработке модели данных. Данная группа работает под эгидой of the Commission for the Management and Application of Geoscience Information (CGI), это новая комиссия of the International Union of Geological Sciences (IUGS). Рабочая группа создала три подгруппы: “Conceptual Model/Interchange”, “Testbed” и “Classification Requirements”.

В 1998 г. в Германии была создана правительственная комиссия IMAGI (Interministerial Committee for Geo Information) для разработки и внедрения Германской национальной пространственной базы данных (Geodateninfrastruktur Deuschland: GDI–DE). Главная цель базы данных — гармонизация и представление необходимых геоданных по запросам через Интернет.

Гармонизация информации предполагает решение ряда задач, совокупность которых можно разделить на следующие группы:

1. Организационные задачи. Предполагают определение источников и потребителей данных, системы получения данных и информированность пользователей.

2. Технические задачи. Предполагают реализацию протоколов и стандартов программными и техническими способами, реализацию доступа к данным.

3. Правовые вопросы. Включают разработку лицензионных соглашений, авторских прав и статусов данных, разделение общей информации, организацию защиты, копирования и защиты интеллектуальной собственности.

4. Экономический и социальный аспект. Включают организацию финансирования различных работ и определение стоимости информации и предоставляемых услуг, определение информационного рынка и стоимости, а также получаемой прибыли и ее распределение.

3. Интеграция информации

Объединение информации (доступ к информационным ресурсам) для решения текущих задач (моделирования), рис. 2. Интеграция неизбежно приводит к увеличению объемов данных. Как правило, обусловлена необходимостью оперировать огромными массивами данных в реальном или близком к реальному масштабу времени. Интеграция осуществляется в интересах решения относительно узкого круга задач. Примерами интеграции информации для ГИС могут быть различные форматы, такие как S 57, VPF и ряд других специализированных форматов. С помощью данных форматов информация представляется в определенном виде, в виде структурированных массивов данных. Назначение таких массивов данных — решение определенного круга задач. Например, данные в формате S 57 предназначены для обеспечения навигационной безопасности плавания в заданном районе. Данные в формате SXF обеспечивают решение топографических задач на территории Российской Федерации. В настоящий момент времени наблюдается тенденция разработки сложных, распределенных массивов данных на основе XML–технологии.

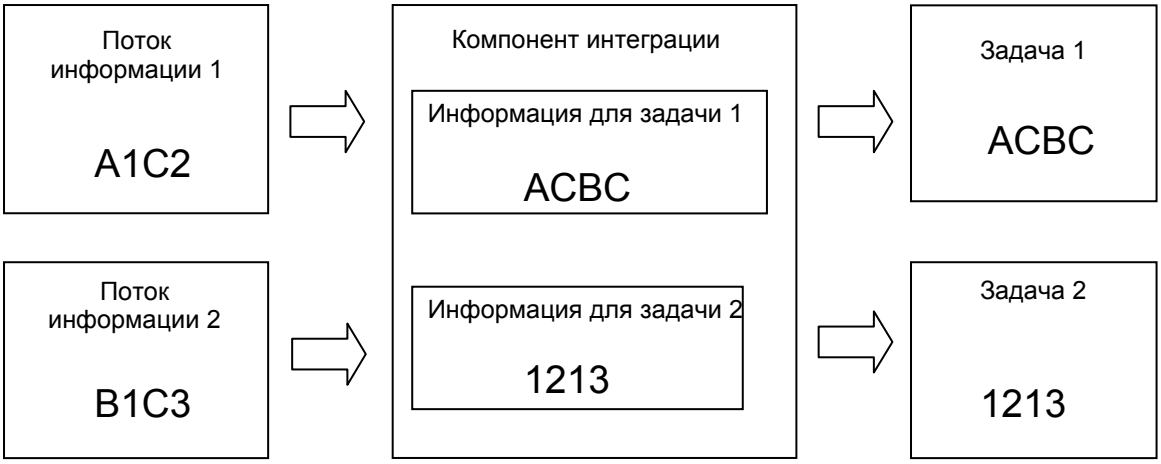

Рис. 2. Интеграция информации.

Доступ к данным осуществляется с использованием различных механизмов и зависит от ряда факторов:

- требуемой скорости обработки данных (реальное время или может быть определенная задержка);
- необходимости параллельной обработки и/или визуализации большого числа данных.

В зависимости от выше перечисленных факторов доступ обеспечивается напрямую в том формате, в котором эти данные хранятся. Но зачастую требуется промежуточное преобразование данных. Такая необходимость появляется, как правило, в системах визуализации ГИС–данных. Это обусловлено техническими ограничениями графических станций и производительностью сети и/или процессоров. Например, данные в формате VPF могут занимать (в зави-

симости от масштаба) сотни гигабайт информации и более. Никакая умная навигация не способна обеспечить работу с таким огромным массивом данных в реальном масштабе времени. Но тем не менее интегрированные по данному формату данные несравнимо удобнее для дальнейшей обработки, чем просто гармонизированные данные, находящиеся в различных источниках с различной скоростью и дисциплиной доступа.

Отличительной чертой интеграции информации является то, что результат направлен на решение определенного класса задач.

Интеграция информации предполагает определение определенной модели данных. Пример теоретической модели данных формата S 57 показан на рис. 3. Данная модель является достаточно сложной, если учесть дальнейшую детализацию векторного представления данных. Сложность определяется тем фактом, что для навигационных целей требуются *взаимосвязанные данные* различного рода, такие как изобаты, навигационные знаки, фарватеры и т.д. Изменение или обновление одних данных может оказывать влияние на другие, взаимосвязанные данные.

Кроме теоретической модели разработан формат хранения данных на машинном носителе, который предполагает один или несколько взаимосвязанных файлов.

Следует заметить, что интеграция данных отнюдь не предполагает физического объединения информации в одном месте, например у локального пользователя. Все зависит от поставленной задачи и конкретных условий. Например, что касается любого формата, обеспечивающего навигационную безопасность плавания, то на конкретном судне должна быть полная информация по тому району плавания, где он находится или предполагает быть. И совсем другой вопрос обеспечения своевременной корректуры данной информации. При решении, например, исследовательских задач информация физически может быть распределенной в некоторой сети, локальной или глобальной.

Интеграция информации предполагает решение ряда задач:

1. Организационные задачи: определение источников и потребителей данных, системы получения данных, их взаимосвязь и система обновлений.

2. Технические задачи. Реализация и развитие форматов данных как на транспортном, так и на интерфейсном уровне. Разработка и внедрение системы производства, распределения, защиты и корректуры данных в заданном формате.

3. Правовые вопросы. Разработка лицензионных соглашений, авторских прав и статусов данных, разделение общей информации, организация защиты, копирования и защиты интеллектуальной собственности.

4. Экономический и социальный аспекты. Организация финансирования различных работ и определение стоимости информации и предоставляемых услуг. Определение информационного рынка и стоимости, а также получаемой прибыли и ее распределение.

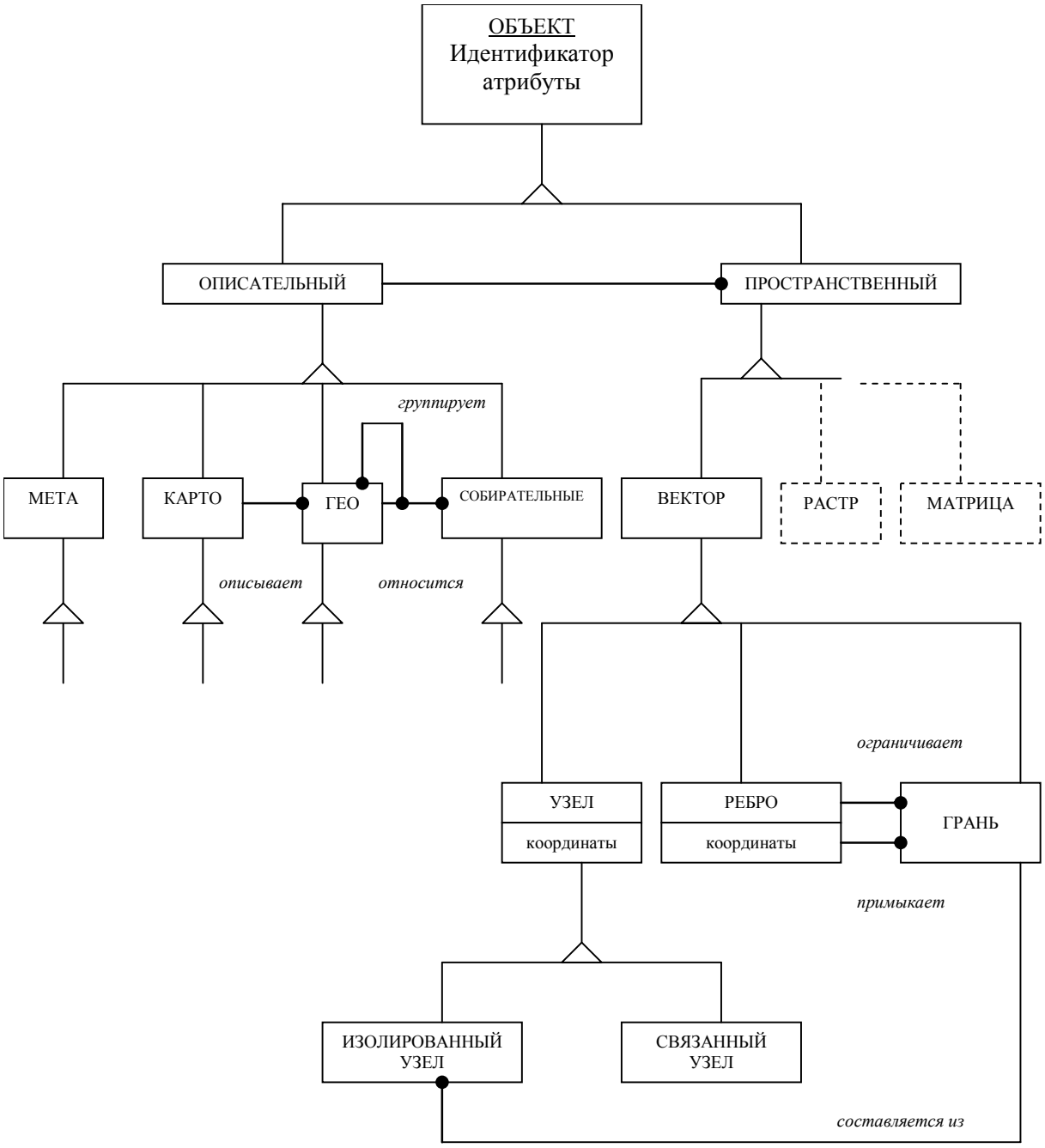

Рис. 3. Теоретическая модель данных формата S 57.

4. Слияние информации

Получение нового качества информации (уменьшение объема информации). Является наиболее сложным этапом преобразования данных. Данное название ассоциируется с известной областью исследований, которая насчитывает не одно десятилетие. Качественным шагом в развитии данного направления явилась разработка модели слияния данных (СД). Данная модель известна как Joint Directors of Laboratories (JDL) Data Fusion Model [4]. В [5] отмечается, что «слияние» может рассматриваться в различных контекстах:

- программное обеспечение (Cold Fusion, e Business);

- физика (медицина, ядерный синтез);
- комбинирование (объединение — соединение различных элементов в некоторое объединение; интеграция — составление некоторого целого из составных частей);
- знания (слияние данных, слияние данных обнаружителей и слияние информации);

В [5] понятия «данные» и «информация» разделены. Слияние данных — это организованное комбинирование в интересах анализа и принятия решения, а слияние информации — это комбинирование данных для получения знания. В [3] СД определяется как процесс соединения данных от *различных источников*. Цель СД определяется как получение информации более *высокого качества*. При этом понятие *высокое качество* зависит от области применения. Можно заметить, что большинство исследований СД действительно определяет главную цель СД как повышение качества информации (данных).

Для большинства современных ГИС–приложений проблема *высокого качества* данных превратилась в последовательность корректно сформулированных и поставленных задач, имеющих различные варианты решений и обеспечивающих *высокое качество* данных для конкретно поставленной задачи.

В настоящий момент времени проблемой является не столько *высокое качество* данных, а, как правило, изменение качества данных.

Изменение качества данных требует серьезной аналитической проработки предметной области. И в данном контексте мы и будем понимать слияние данных (информации) для ГИС–технологий. Суть слияния (сплава, синтеза) информации показана на рис. 4. Для примера, приведем схему слияния информации для систем мониторинга различного назначения, рис. 4. На данном рисунке показано изменение качества информации снизу вверх по иерархии. Если учесть, что системы мониторинга являются сложными и распределенными в пространстве, то становится очевидной идея слияния данных. Без данного механизма подобная сложная система просто не сможет работать.

Отличительной особенностью процесса слияния информации является получение нового качества информации и сокращение объема.

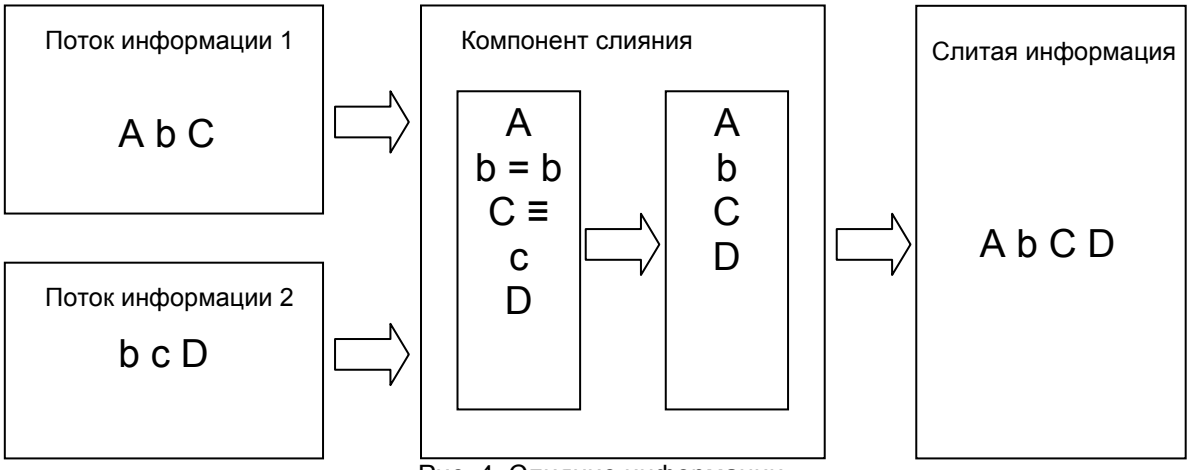

Рис. 4. Слияние информации.

Обозначенные на рис. 4 уровни представляют собой качественные скачки в представлении информации. Данная схема применима практически для любой системы мониторинга, а возможно, и не только для данного класса систем. В данном случае иллюстрируется гегелевский принцип преобразования количе-

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ства в качество. Существует, однако, нюанс, заключающийся в том, что не существует универсальных механизмов для организации подобных качественных преобразований. Это целая система специальных исследований, включающая целый ряд научных направлений.

Дадим более детальные комментарии к рис. 4.

На нулевом (первом) уровне осуществляется выделение полезного сигнала на фоне помех системой или средством измерения физических полей.

На первом уровне принимается решение об обнаружении сигнала определенного класса. Данные уровни отличаются как исходной информацией, так и математическими методами. На нулевом уровне мы имеем дело с классической теорией обнаружения сигналов, а на первом — с методами классификации или распознавания. Иногда на первом уровне появляется еще один — трассовый анализ, который обладает достаточной автономностью как по методам исследования, так и по областям применения.

На втором уровне происходит анализ возможных ситуаций, которые создаются действиями или бездействием определенных объектов. Это — очередной качественный скачок, как в смысле предмета исследования, так и применяемых методов исследования.

Анализ возможных ситуаций, как правило, не является самоцелью и следующий этап — это оценка потенциальных угроз, которые создает текущая ситуация.

Выделенные выше уровни обработки информации являются взаимосвязанными и взаимозависимыми, как показано на рис. 5.

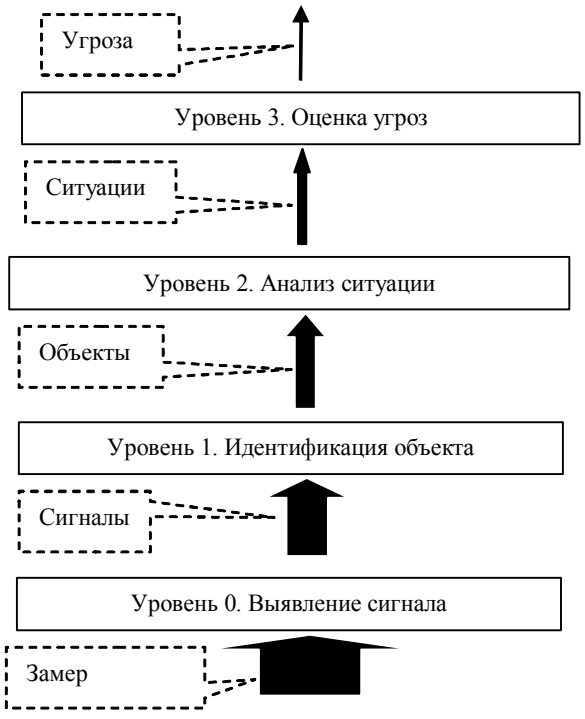

Рис. 5. Слияние информации в системах мониторинга.

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

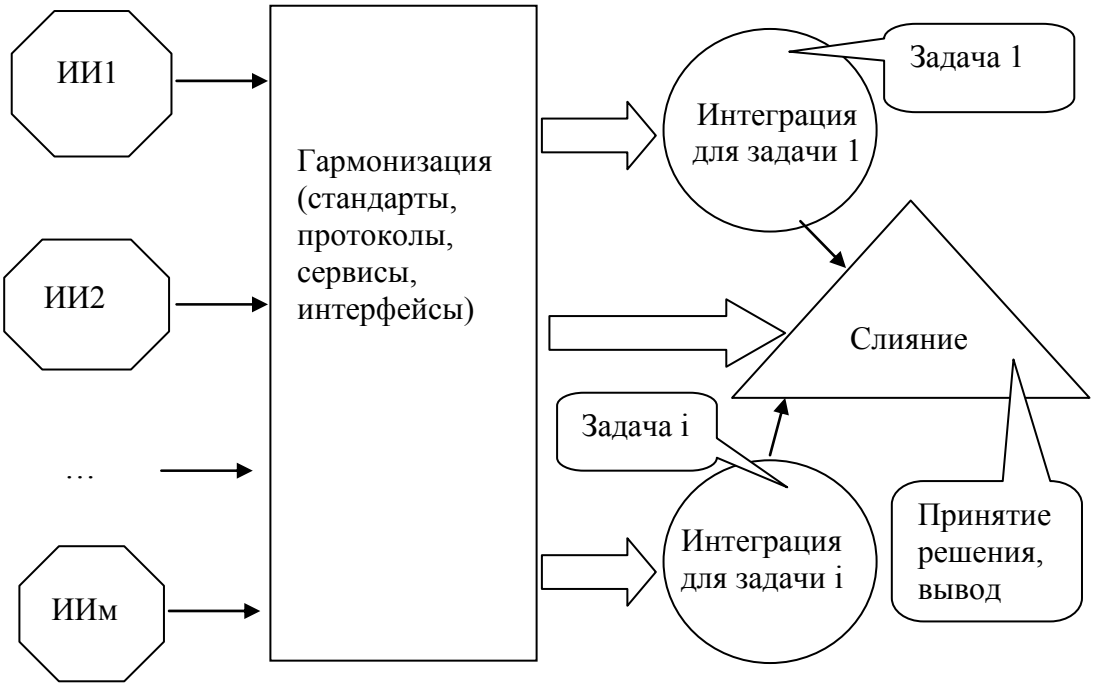

Рис. 6. Взаимосвязь уровней обработки информации.

Слияние информации предполагает некоторую последовательность действий.

1. Организационные задачи: определение источников и потребителей данных, системы получения данных, их взаимосвязь и система обновлений.

2. Технические задачи. Реализация и развитие форматов данных как на транспортном, так и на интерфейсном уровне. Разработка и внедрение системы производства, распределения, защиты и корректуры данных в заданном формате.

3. Правовые вопросы. Разработка лицензионных соглашений, авторских прав и статусов данных, разделение общей информации, организация защиты, копирования и защиты интеллектуальной собственности.

4. Экономический и социальный аспекты. Организация финансирования различных работ и определение стоимости информации и предоставляемых услуг. Определение информационного рынка и стоимости, а также получаемой прибыли и ее распределение.

5. Заключение

Рассмотренные в данной статье вопросы обработки информации в ГИС и ГИС–приложениях показывают, что проблема в настоящий момент времени выходит достаточно далеко за рамки традиционных исследований ГИС и их практических приложений. Любая попытка внедрения ГИС в реальные системы, как правило, приводит к возникновению выше перечисленных задач: гармонизации, интеграции и слияния данных.

Возникает острая необходимость в теоретическом и практическом исследовании различных уровней обработки информации для ГИС и их приложений. Некоторые проблемы уже решаются непосредственно или опосредовано. Такие технологии, как вэб–сервисы, концепция сервис–ориентированных архитектур непосредственно применимы для того, чтобы реализовать гармонизацию информации. Различные расширения технологии языков разметки, например

GML, также могут и должны быть использованы для решения вопросов интеграции и гармонизации данных.

Проблема интеграции данных исторически решалась разработкой специализированных форматов или наборов данных. Наиболее известные из них S57, VPF, SXF, а также различные Shape форматы. Но как показывает опыт применения ГИС, всегда возникает необходимость использования как минимум данных из двух и более форматов (т.е. глобальных источников).

Слияние данных в ГИС — наиболее сложная область исследований и технологических решений. Особенность данного уровня — еще более узкая ориентация по сравнению с интеграцией на конкретного пользователя. Плюс к этому — необходимость использования достаточно сложных математических подходов и моделей.

Литература

1. Webopedia [Электронный ресурс] // <http://www.webopedia.com/TERM/d/data.html> (по состоянию на 01.02.2008).
2. *Asch, K.* An International Initiative for Data Harmonization in Geology / *K. Asch, B. Brodaric* // 10th EC–GI&GIS Workshop: ESDI: The State of the Art, Warsaw, Poland, 23–25 June 2004. — P. 9–15.
3. *Valet, L.* A statistical overview of Recent Literature in Information Fusion / *L. Valet, G. Mauris, P.* A *Bolon* // 3rd International Conference on Information Fusion: process., France. 10–13 July 2001. — P. 532–536.
4. *White, F. E.* A Model for Data Fusion / *F. E. White* // 1st National Symposium on Sensor Fusion: process., vol.2, 1988. — P. 20–26.
5. *Blasch, E.* Fundamentals of Information Fusion and Applications / *E. Blasch* // Tutorial, TD2, Fusion 2002. — 230 p.
6. *Llinas, J.* Revising the JDL Data Fusion Model II / *J. Llinas* // 9th International Conference on Information Fusion: process., Philadelphia, USA, 25 - 29 July 2005. — P. 230–238.