

Унификация средств ведения архива данных ДЗЗ

В.П. Саворский, Ю.Г. Тищенко

*Учреждение Российской академии наук Институт радиотехники и электроники
им. В.А. Котельникова РАН, Фрязинский филиал
141190 Фрязино, Московская обл., площадь им. академика Б.А. Введенского, д.1
E-mail: (savor,tishchen)@ire.rssi.ru*

Рассматривается один из перспективных подходов к решению проблем, связанных с повышением эффективности усвоения разнородной информации в информационных системах - унификация процедур обмена и хранения данных. Сформулированы требования к унифицированным программно-информационным средствам архива, разработаны унифицированные процедуры автоматизированного ведения архива данных.

Ключевые слова: информационная система, унификация, архив данных, буфер, директория, автоматизированная система.

Введение

Комплексный анализ данных, получаемых в процессе многолетних исследований природных объектов средствами дистанционного зондирования Земли (ДЗЗ), в общем случае, предполагает включение в процедуры обработки данные из различных источников. Существенную часть этих данных составляет первичная исходная информация, т.е. данные, полученные в результате дистанционных либо вспомогательных контактных наблюдений, а также получаемые на их основе тематические продукты. Эти данные, в общем случае, различаются как по составу включаемых в них параметров, так и по форме их представления. Основной причиной этих различий является то, что первичные данные и продукты ДЗЗ формируются, как правило, в рамках независимых проектов, которые проводятся независимыми группами исследователей с помощью различающихся средств наблюдения, регистрации и обработки данных. Следует отметить, что даже одна и та же группа исследователей с течением времени в силу различных причин изменяет состав и форматы генерируемых в ее проектах наборов данных.

Разброс в представлении исходной информации, используемой в системах комплексного анализа данных ДЗЗ, существенным образом усложняет их развитие и эксплуатацию. Наиболее остро это проявляется на таких этапах развития информационных систем ДЗЗ (ИС ДЗЗ), когда требуется их быстрая и надежная перенастройка при появлении новых источников данных. В частности, отсутствие единых подходов к процедурам и формам обмена и хранения данных приводит к существенному удорожанию работ по модернизации систем обработки и анализа. Важно отметить, что это приводит также к снижению надежности и устойчивости работы ИС ДЗЗ, т.к. каждый раз при разработке и введении новых элементов в уже существующую систему, зачастую, весьма вероятны потери контроля над необходимыми для ее функционирования ресурсами.

Одним из перспективных подходов к решению проблем, связанных с повышением эффективности усвоения разнородной информации в ИС ДЗЗ, является унификация процедур обмена и хранения данных. **Целью данной работы** и является разработка программно-информационных средств ввода данных в архив, поиска в архиве необходимых данных и их

извлечения из архива в таком виде, который позволяет полностью автоматизировать указанные процедуры. При таком подходе эффективность унификации оценивается в соответствии с тем, способствует ли предлагаемая унификация представления данных и организации процедур обмена автоматизации комплексного анализа данных ДЗЗ или нет.

1. Базовые требования к унифицированным средствам архива

Для достижения поставленной цели необходимо, прежде всего, сформулировать требования к процедурам унификации. Анализ показывает, что унификация не может быть достигнута, если предлагаемые для достижения поставленной цели решения, будут существенным образом зависеть от:

- используемых в ИС ДЗЗ программно-аппаратных платформ,
- используемых в ИС ДЗЗ средств архивирования и обмена данными,
- форматов хранения или представления данных.

Независимость унифицированных решений от программно-аппаратных платформ и используемых средств хранения и обмена данными означает, что целевые решения должны быть переносимы с платформы на платформу без существенных затрат на изменение программного кода. Переносимость, таким образом, означает, что используемые решения должны базироваться на таких средствах разработки и поддержки функционирования программного обеспечения, которые присутствуют на всех широко распространенных программно-аппаратных платформах. В частности, это означает, например, что целесообразным представляется для описания процедур обмена данными (файлами) использовать процедурный язык Perl, для ведения каталогов – СУБД Oracle или MySQL, для хранения данных – накопители со SCSI или FC (SAN) интерфейсами.

Указанные требования к унифицированным процедурам ИС ДЗЗ означают также, что унификация в предлагаемом нами подходе включает только те подсистемы ИС ДЗЗ, в которых нет необходимости интерпретировать (а тем более необратимо изменять) смысловое содержание исходных данных ДЗЗ и/или получаемых на их основе продуктов. Это означает, что унифицированные процедуры производят манипуляции с файлами данных ДЗЗ, не изменяя необратимо состав этих файлов. Таким образом, объектами нашего анализа являются файлы с данными, с которыми унифицированные процедуры должны производить такие манипуляции, которые не приведут к потере их содержания, описываемого в виде структурированного набора атрибутов. При этом структурированный набор атрибутов (по существу, паспорт файла данных) может быть описан как в виде внутреннего объекта, размещенного внутри самого файла данных, так и вида внешнего объекта, например, в виде файла с метаданными (метафайла) или в виде таблиц базы данных.

2. Основные задачи

Учет требований и ограничений, сформулированных в п.1, позволяет сформулировать основные задачи, решение которых в совокупности обеспечит унификацию средств ведения архива данных ДЗЗ. Такими принципиальными задачами являются

- унификация средств загрузки архива данными ДЗЗ,

- унификация средств тиражирования данных, хранимых в архиве ДЗЗ,
- универсализация средств контроля состояния буферов обмена и хранения данных.

Решение задач по унификации средств загрузки и извлечения данных с архивных средств, а также контроля состояния средств обмена являются основными задачами унификации. При этом надо учитывать то, что в общем случае в конфигурацию архивной системы входят, как правило, несколько серверов и рабочих станций. Кроме того, само функционирование ИС ДЗЗ предполагает возможность и необходимость обмена массивами данных с удаленными клиентами. Поэтому при решении указанных задач важно выполнить следующее **основное требование по обмену данными**: создаваемые унифицированные средства должны включать в себя процедуры автоматического обмена данными с удаленными партнерами – как поставщиками, так и потребителями.

3. Принципы решения

Решение сформулированных в п.2. задач базируется на следующих принципиальных подходах:

- использование программных средств, не привязанных к конкретным программно-аппаратным платформам – Perl, Oracle,
- создание и эксплуатация буферов хранения и обмена данными с единой унифицированной самоописываемой структурой, включающей
 - уникальное имя буфера,
 - директории с группами данных (подразделами архива), характеризующимися групповым именем в виде единого шаблона имени директории,
 - файлы с данными, характеризующимися уникальным признаком (расширителем),
 - файл описания, или файл готовности, состава директории (подразделами архива), имеющий то же, что и директория имя и уникальный расширитель, этот файл, по минимуму, содержит список файлов директории и их размеры и предназначен как для контроля состава директории, так и для подтверждения завершения записи данных в директорию (для этого файл должен быть создан и записан в эту директорию после записи всех файлов с данными),
- использование универсальных средств обмена и контроля содержания буферов – команды ftp и созданные на их основе утилиты потоковой обработки,
- использование единого универсального регистра файлов архива, базирующегося на СУБД с возможностями удаленного доступа – Oracle, MySQL.

4. Структура буферов архива

На рис.1 в качестве примера приведена структура буферов архива данных прибора MODIS (коллекция IRE_TERRA_MODIS) и прибора КМСС (коллекция IRE_METEOR_KMSS), установленных на космических аппаратах TERRA и «Метеор-М» №1 соответственно.

Структура буферов определяется, прежде всего, их назначением (APPOINTMENT). Так, на приведенной на рис.1 схеме можно выделить 5 типов буферов системы хранения и обмена данными, определяемых их назначением: 1) долговременный архив, 2) опера-

тивный архив, 3) входной буфер системы архивации (может быть совмещен с оперативным архивом) 4) выходной буфер системы обработки, 5) входной буфер системы выполнения заказов (СВЗ). В каждом из этих буферов создаются дочерние буфера для хранения в них отдельных коллекций данных (DATACOLLECTION). На рис.1. это коллекции IRE_TERRA_MODIS и IRE_METEOR_KMSS.

Архив состоит из 2-х ключевых разделов: 1) оперативный архив, 2) долговременный архив. Основным разделом архива является долговременный архив. Именно в этом разделе обеспечивается полная безусловная сохранность всех данных ИС ДЗЗ, которые предназначены к долговременному хранению. Оперативный архив играет роль промежуточного буфера между долговременным архивом и потребителями/поставщиками данных. Его основное назначение – обеспечивать быстрый доступ к актуальным и востребованным данным. Показанный на рис.1 входной буфер системы архивации предназначен для проведения входного контроля и переупаковки данных перед включением их в архив. Он может быть совмещен с оперативным архивом в том случае, когда проверка (например, при поступлении данных, сгенерированных в самой ИС ДЗЗ) и переупаковка не требуется.

Выходной буфер системы обработки служит для сбора данных на выходе системы обработки перед передачей их на хранение в архив, а входной буфер системы выполнения заказов – для временного хранения заказанных данных перед передачей внешним потребителям.

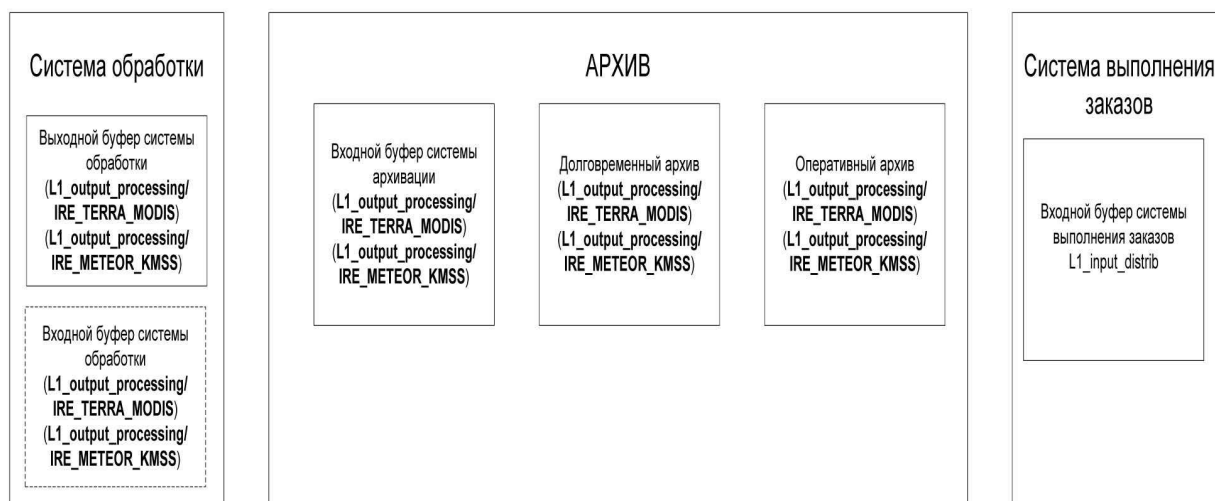


Рис.1. Структура буферов архива, коллекции IRE_TERRA_MODIS, IRE_METEOR_KMSS

5. Регистр архива

Регистр архива хранится в 2-х таблицах базы данных Oracle:

- регистр директорий – DIR_REGISTRY,
- регистр файлов – FIL_REGISTRY.

Каждая из этих таблиц включает следующие атрибуты описания объектов архивной системы (фалов и директорий):

- идентификатор,
- имя,
- размер,
- даты создания/регистрации/изменения/ замещения/удаления/,

- ссылку на замещающий объект (файл или директорию, соответственно).

Регистр файлов хранит также ссылку на родительскую директорию, а регистр директорий – также указания на назначение директории APPOINTMENT и имя хранимой в ней коллекции DATACOLLECTION.

6. Унифицированные средства конфигурирования

Параметры конфигурация всех буферов хранятся в регистре буферов – таблице BUFFER_DESCR. Каждый буфер однозначно характеризуется своим назначением (APPOINTMENT) и именем хранимой в нем коллекции данных (DATACOLLECTION). Полная конфигурация буфера включает атрибуты, описывающие:

- расположение буфера,
- условия доступа,
- групповые имена файлов данных и файлов описаний,
- имена буферов – источников данных для данного буфера (для организации потоковой автоматической подачи данных),
- условия хранения,
- средства хранения.

7. Унифицированная система контроля состояния архива

Единообразное описание и унификация структуры буферов позволяют создать и использовать унифицированную систему контроля состояния всех буферов архива, а следовательно, и архива в целом. Основными структурными элементами этой системы являются:

- подсистема просмотра буферов архива (утилита `buffer_register3`) предназначена для формирования задания (списка буферов) для утилиты `dir_registry_admin`,
- подсистема регистрации состояния директорий, содержащихся в контролируемом буфере (утилиты `dir_registry_admin`, `family_short_list`, `dir_registry_filcheck`) предназначена для регистрации актуального состояния директорий контролируемого буфера,
- подсистема регистрации состояния файлов, содержащихся в контролируемой директории (утилита `fil_registry`) предназначена для регистрации актуального состояния файлов заданной директории в контролируемом буфере,
- подсистема автоматической коррекции записей регистра (утилиты `no_orphan`, `no_blind`) обеспечивает поддержку целостности регистра как при работе в режиме замещения (`no_orphan`), так и для обеспечения соответствия регистровых записей актуальному состоянию регистров архива (`no_blind`),
- регистр архива данных (таблицы `BUFFER_DESCR`, `DIR_REGISTRY`, `FIL_REGISTRY` под управлением СУБД Oracle) предназначен для регистрации атрибутов описания и состояния составляющих архив структурных единиц (буферов, директорий и отдельных файлов),
- сводки допустимых имен (таблицы `DATA_SOURCE_TYPE_LIST`, `DIR_APPOINT`

³ Все упоминаемые в работе утилиты реализованы в виде скриптов на языке Perl

MENT_LIST, DIR_STATUS_LIST, FIL_STATUS_LIST, STORAGE_MODE_LIST, STORAGE_TYPE_LIST под управлением СУБД Oracle) задают унифицированное описание регистрируемых объектов данных.

Функционирование системы контроля представлено на рис. 2.

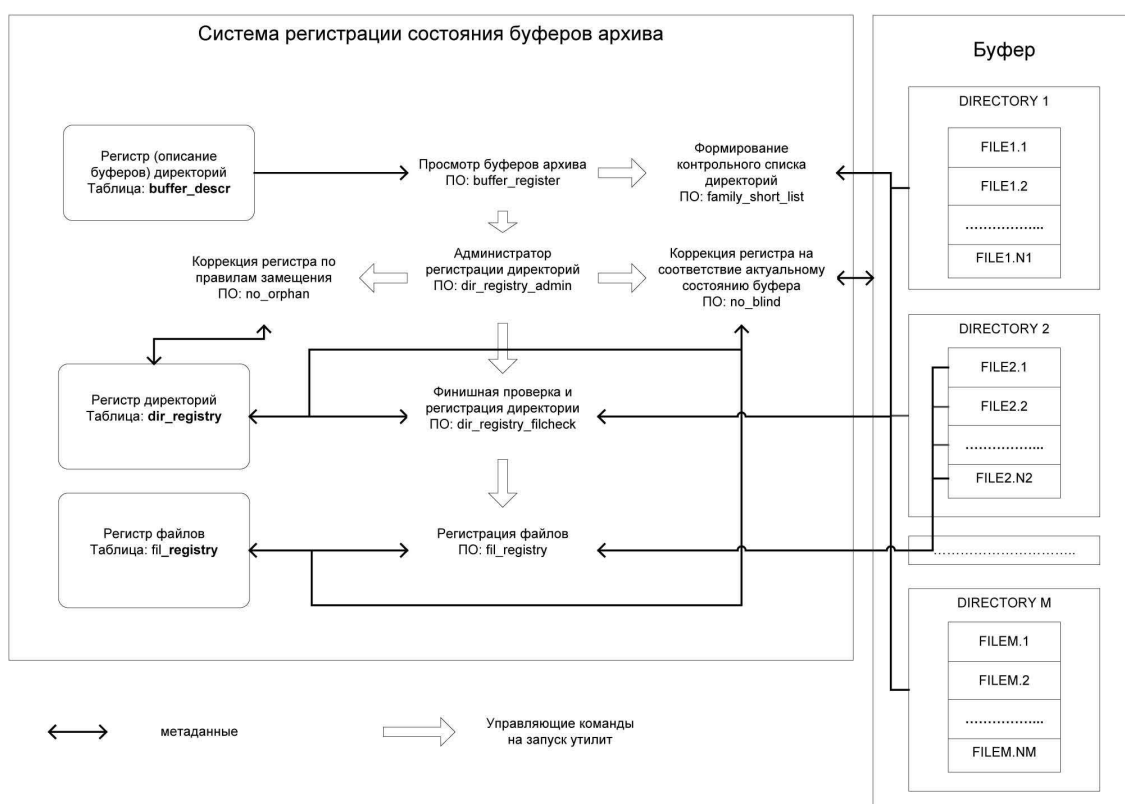


Рис. 2. Функциональная схема системы контроля буферов архива

8. Унифицированная автоматизированная системы ввода данных в архив

Основными структурными элементами автоматизированной системы ввода данных в архив являются:

- администратор процедур загрузки данных в архив (утилита archive_ingest), назначение которого заключается в формировании заданий на загрузку данных в буфера архива и запуск (в соответствии со сформированным заданием) утилит download_ready_dir, fix_dir_diffs и скрипта upld_auto,
- подсистема загрузки данных на дисковые массивы (утилиты download_ready_dir, fix_dir_diffs), назначение которой заключается в загрузке полных директорий в буфер в соответствии с заданием, сформированным администратором (archive_ingest) утилитой download_ready_dir, либо перемещении отдельных файлов, указанных в задании, утилитой fix_dir_diffs, базовые процедуры перемещения обеспечивает утилита ftp_common_cmd,
- подсистема загрузки данных на роботизированные ленточные накопители (утилиты upld_auto, upld2lib), назначение которой заключается в загрузке полных директорий на роботизированный ленточный накопитель в соответствии с заданием, сформированным администратором (archive_ingest), который запускает менеджер загрузки upld_auto, обеспечивающей активизацию политики загрузки и запуск самой утилиты загрузки upld2lib,

- подсистема выборки данных из архива (утилиты `robot_file_restore`, `oper_restore`, `forward2distrib`), назначение которой заключается в обеспечении выборки данных из долговременного и оперативного архивов,
- регистр архива данных (таблицы `BUFFER_DESCR`, `DIR_REGISTRY`, `FIL_REGISTRY` под управлением СУБД Oracle) предназначен для регистрации атрибутов описания и состояния составляющих архив структурных единиц (буферов, директорий и отдельных файлов).

Функционирование автоматизированной системы ввода данных в архив представлено на рис. 3. Ввод данных в буфер инициируется автоматически по событию, которым является наличие в буфере-источнике данных, которые еще не зарегистрированы в буфере-приемнике. Эти данные формирует система контроля состояния буферов, а система ввода использует актуальные данные о состоянии для вынесения решения о начале переноса данных при их наличии.

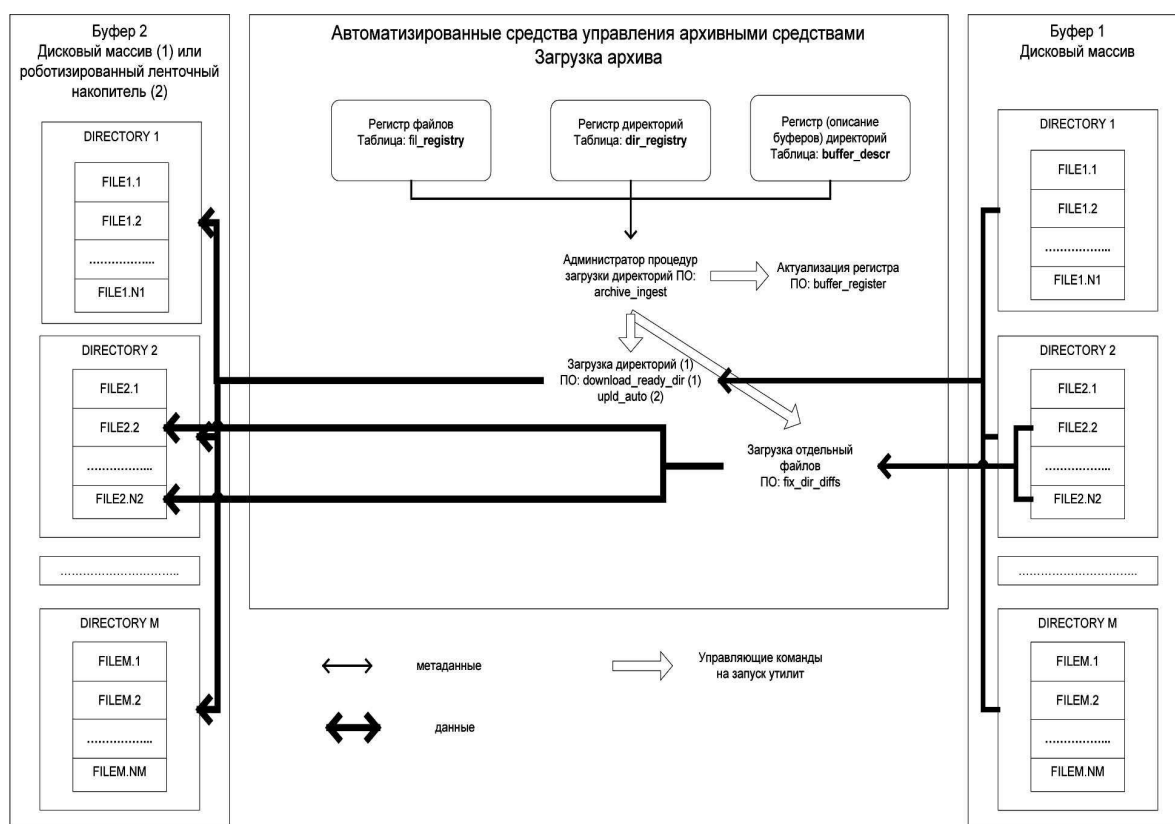


Рис. 3. Функциональная схема автоматизированной системы ввода данных в архив

9. Унифицированная автоматизированная системы извлечения данных из архива

Функциональная схема автоматизированной системы извлечения данных из архива представлена на рис. 4. Основными структурными элементами системы являются:

- подсистема мониторинга заказов СВЗ (утилита `check_new_order`), назначение которой заключается в обеспечении постоянного, т.е. круглосуточного, контроля за появлением новых заказов в таблице заказов, кроме того, эта подсистема обеспечивает регистрацию принятого заказа и отметок об его исполнении как в регистре СВЗ, так и в регистре заказов архива,

- подсистема автоматического переноса заказанных данных из долговременного архива в оперативный (утилита `oper_restore`), назначение которой заключается в извлечении данных из долговременного архива и запись их в оперативный, она применяется в автоматизированной системе тиражирования в том случае, когда заказанный файл не найден в оперативном архиве,
- подсистема автоматического переноса данных из оперативного архива в буфер СВЗ (утилита `forward2distrib`), назначение которой – перенос данных из оперативного архива в выделенный для распределения данных буфер системы выполнения заказов,
- регистр выданных из архива данных (таблица `order_registry` под управлением СУБД Oracle), предназначенная для ведения списка выдачи данных из архива.

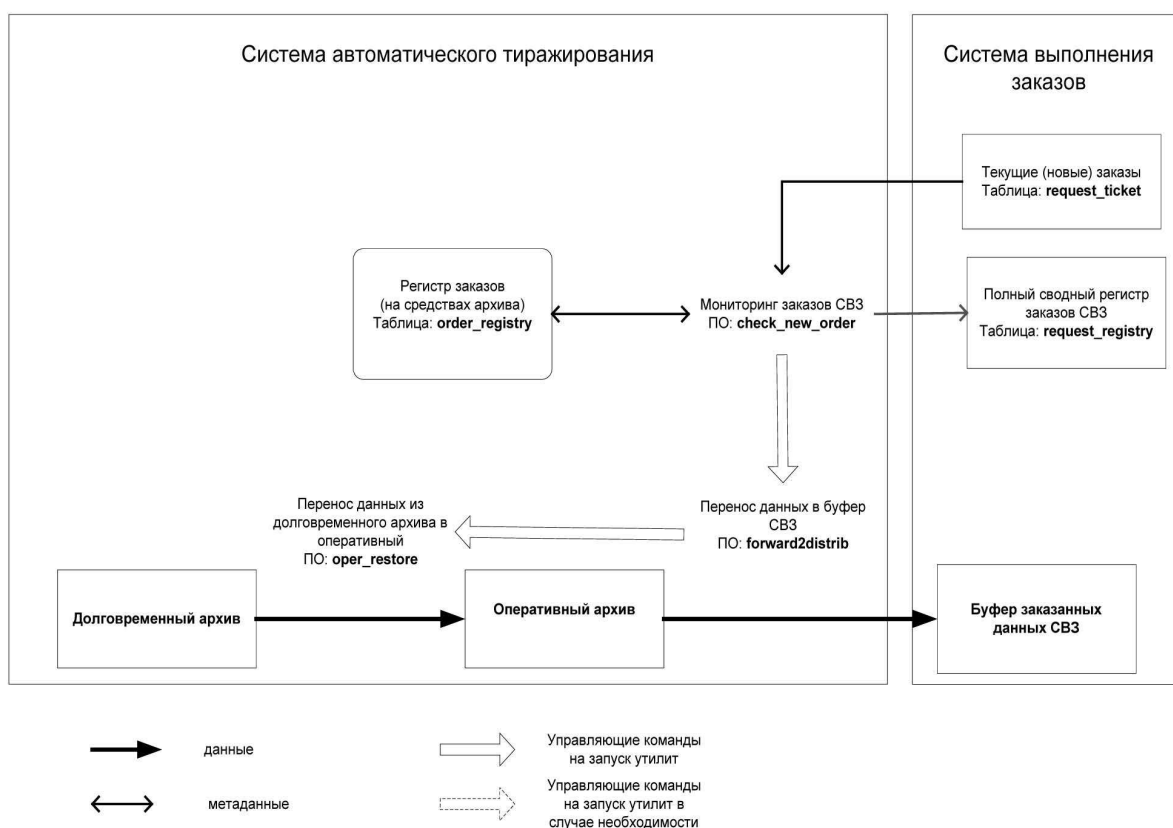


Рис. 4. Функциональная схема автоматизированной системы извлечения данных из архива

Заключение

В работе проведен анализ одного из перспективных подходов к решению проблем, связанных с повышением эффективности усвоения разнородной информации в информационных системах – унификация процедур обмена и хранения данных. Сформулированы требования к унифицированным программно-информационным средствам архива: для описания процедур обмена данными (файлами) целесообразно использовать процедурный язык Perl, для ведения каталогов – СУБД Oracle или MySQL, для хранения данных – накопители со SCSI или FC (SAN) интерфейсами. Разработаны унифицированные процедуры автоматизированного ведения архива данных: конфигурирования, систем контроля состояния архива, ввода данных в архив, извлечения данных из архива.

Software Unification of the Earth Remote Sensing Data Archive Management

V.P. Savorskiy, Yu.G. Tishchenko

*Institute of Radioengineering and Electronics, Russian Academy of Sciences,
Fryazino, Moscow Region
E-mail: (savor,tishchen)@ire.rssi.ru*

One of the perspective approaches to solve of problems, concerning of efficiency improve of difference information assimilation in information systems is discussed, namely unification of data exchange and data storage procedures. The requirements to archive unification software are formulated. Unified procedures of software of the Earth remotes sensing data archive automated management are worked out.

Keywords: information system, unification, data archive, buffer, directory, automated system.