РОССИЙСКАЯ АКАДЕМИЯ НАУК
СПЕЦИАЛЬНАЯ АСТРОФИЗИЧЕСКАЯ ОБСЕРВАТОРИЯ

# ПРЕПРИНТ 219

Желенкова О.П., Витковский В.В., Пляскина Т.А.,
Малькова Г.А., Шергин В.С., Марухно А.С.

## Открытый Интернет-доступ к общему архиву наблюдений САО РАН в контексте стандартов виртуальной обсерватории

Нижний Архыз
2007

*УДК 520.52-13(083.8)*

# Открытый Интернет-доступ к общему архиву наблюдений САО РАН в контексте стандартов виртуальной обсерватории

**Желенкова О.П., Витковский В.В., Пляскина Т.А., Малькова Г.А., Шергин В.С., Марухно А.С.**

Специальная астрофизическая обсерватория РАН, п.Нижний Архыз, 369167, Россия

**Аннотация.** В 2003 году Генеральная Ассамблея Международного астрономического союза приняла резолюцию об открытом Интернет-доступе к архивированным наблюдениям, полученным в обсерваториях, которые финансируются из государственного бюджета. Методы работы с данными в виртуальной обсерватории определяются стандартами Международного альянса "Виртуальная обсерватория" (International Alliance Virtual Observatory или IVOA), предъявляющими определенные требования к уровню подготовки ресурсов. К сожалению, не имеется готовых процедур и программного обеспечения для приведения данных к принятым стандартам. В нашем случае, когда архив наблюдений состоит из разнородных цифровых коллекций, накапливаемых в течение 20 лет, такая работа требуют значительных трудозатрат. Реинжениринг архивных данных обсерватории ведется поэтапно, начиная с анализа полноты необходимых параметров в каждом локальном архиве и коррекции, где это возможно, пропущенных или ошибочных параметров в отдельном наблюдательном файле.

Обеспечен открытый доступ из Интернета к наблюдательным данным общего архива обсерватории. Запросы реализуются с помощью информационно-поисковой системы (ИПС), которая размещена на специализированном сервере. Каждый файл с наблюдениями описывается в таблицах базы данных набором из более, чем 60 параметров. Они используются для динамического формирования web-интерфейса, установления соответствия FITS-параметров с параметрами таблиц ИПС, идентификации файлов, определения типа файла и т.п. Часть таблиц схемы базы данных являются справочниками (словарями) и содержат информацию, собранную при анализе данных локальных архивов. Она позволяет интегрировать разнородные данные в одной ИПС. Другая часть таблиц содержит параметры, описывающие каждый файл с наблюдениями, включенный в архив. Таблицы пополняются по мере поступления оптических дисков с новыми данными. В архиве нет жестких ограничений на формат файлов, поэтому добавление новых локальных архивов не вызывает трудностей при соблюдении достаточно простых правил. В статье отражено текущее состояние архивной системы и планы по ее развитию.

## 1. Введение

Современные астрономические телескопы выдают большой объем цифровых данных. В настоящее время исследователю приходится затрачивать значительное время на организацию разнородной информации и обработку наблюдений, поэтому при совместном сотрудничестве в группе, особенно, если в нее входят исследователи из разных учреждений, необходимы соглашения о форматах и понятное всем описание данных или, иначе говоря, стандартизация. Стандарты требуются:

- при форматировании данных; чтобы одна группа с легкостью могла читать и понимать данные, полученные другой группой;
- в семантике; чтобы можно было термин, применяемый одной группой при описании параметров, перевести в используемый другой группой термин, аналогичный по смысловому значению;
- в процедурах обработки; чтобы отдельные этапы можно было выполнять в Интернете с возможностью воспроизведения этапов обработки другими исследователями.

Пятая комиссия Международного астрономического союза, называемая "Documentation and Astronomical Data", занимается вопросами документирования, стандартизации и менеджмента астрономических данных. Астрономические наблюдения не теряют со временем своей научной значимости, но большое количество накопленной информации в обсерваториях пока недоступно исследователям по разным причинам. В 2003 году Генеральная Ассамблея МАС приняла резолюцию об открытом доступе к архивированным данным (Public Access ..., 2003). Наблюдения, которые получены обсерваториями, финансируемыми из государственного бюджета, после определенного периода, когда они являются собственностью заявителей программы, должны быть помещены в архив с открытым Интернет-доступом. На сколько это возможно, данные должны сопровождаться соответствующими описаниями и программными средствами, чтобы можно было извлекать из них научную информацию для последующего анализа.

В конце 20 – начале 21 века активное внедрение современных информационных технологий в астрономические исследования привели сначала к появлению концепции, а затем к созданию инфраструктуры виртуальной обсерватории (ВО). Основная идея ВО состоит в реализации прозрачного доступа к распределенным данным, организации простой и эффективной работы с ними так, как если бы информация находилась на компьютере пользователя (Building the Framework ..., 2004). Координирует и направляет деятельность астрономического сообщества по разработке стандартов ВО созданный в 2002 году Международный альянс "Виртуальная обсерватория" (International Virtual Observatory Alliance или IVOA).

Виртуальная обсерватория оперирует с готовыми для научного анализа каталогами и обзорами. Включение в нее архивов наблюдений является более сложной задачей, требующей разработки стандартов и программного обеспечения для всех этапов проведения наблюдений (от подачи заявок до обработки данных) с тем, чтобы получить единый технологический цикл, конечная цель которого состоит в помещении результата в информационную инфраструктуру виртуальной обсерватории.

## 2. Методы включения астрономических данных в инфраструктуру виртуальной обсерватории

Семантическая сеть (Semantic Web) является частью концепции развития сети Интернет, целью которой является обработка информации, доступной в WWW, с помощью компьютеров. Основной акцент в в этом случае делается на работе с метаданными, которые однозначно описывают свойства и содержание ресурсов WWW, в отличие от используемого в настоящее время анализа текстовых документов. В семантической сети предполагается повсеместное ис-

пользование универсальных идентификаторов ресурсов, а также онтологий и языков описания метаданных.

Эта концепция принята и развивается Консорциумом W3 (http://www.3w.org), определяющим стандарты и спецификации Интернета. Для ее внедрения предполагается создание сети документов с метаданными, описывающими ресурсы WWW, и существующей параллельно с ними. Тогда как сами ресурсы предназначены для восприятия человеком, метаданные используются машинами (поисковыми роботами и другими интеллектуальными агентами) для автоматического поиска информации и распознавания свойств этих ресурсов.

Виртуальная обсерватория предназначается для применения этого подхода к астрономическим данным, чтобы повысить эффективность работы с ресурсами в сети. На конференции IVOA в 2003 году для создания информационной инфраструктуры виртуальной обсерватории (Williams et al., 2004) были определены шесть основных направлений в разработке стандартов:

• Регистры (Registry). Регистры - это базы данных, в которых собрана информация об астрономических ресурсах и программных сервисах, имеющихся в Интернете. Они обеспечивают запросы по поиску информации, публикуют (делают доступными для обнаружения в Интернете программными средствами) и отслеживают появление новых ресурсов (Plane et al., 2004);

• Модель астрономических данных (Data model). В качестве семантического стандарта для астрономии необходима доменная модель данных, которая позволит разрабатывать программное обеспечение, работающее с множеством вариантов представления данных без модификации структур данных и самих программ;

• Семантический описатель или дескриптор содержимого параметра (Uniform Content Descriptors или UCDs). UCD используется для установления связи между наименованиями параметров и астрономическими понятиями (Derriere et al., 2004). IVOA поддерживает и контролирует словарь дескрипторов по всем разделам астрономии;

• Доступ к данным (Data Access Layer или DAL). DAL включает стандарты, описывающие механизм доступа к распределенным астрономическим данным, и программные средства, обеспечивающие такой доступ;

• Язык запросов для виртуальной обсерватории (VO Query Language). Хотя SQL (Structured Query Language) можно использовать для запросов к большинству современных астрономических баз данных, но астрономическая специфика требует расширения возможностей языка запросов (Ysuda et al., 2004);

• Программные сервисы, для обеспечения распределенных вычислений в Интернете (Grid & Web Services). В настоящее время виртуальная обсерватория – это набор web-сервисов. Так называются программы, которые работают на разных компьютерах в сети и взаимодействуют между собой посредством стандартов WWW. Дальнейшие развитие инфраструктуры - реализация методов асинхронных сообщений, авторизация подписью и управление потоками работ, что реализуется посредством информационных технологий, объединяемых общим названием Grid. Совместная работа сервисов требует наличия общей памяти, где происходит обмен данными. Такая область память называется VOSpace. Эта виртуальная для пользователя, видимая в Интернете память выделяется автоматическим процессом для обмена между задачами;

• Формат данных виртуальной обсерватории (VOTable). Этот формат используется в сервисах, совместимых с протоколами IVOA, для представления результатов запросов (Ochsenbein et al., 2004). Основой VOTable является индустриальный стандарт XML и опыт разработок астрономических форматов FITS (Wells et al., 1981) и CDS Astrores (Ochsenbein et al., 2000).

## 2.1. О стандартах International Virtual Observatory Alliance

Разработкой стандартов для интерфейсов, протоколов, спецификаций занимаются несколько рабочих групп альянса IVOA.

### *2.1.1. Описание астрономических ресурсов*

Рабочая группа Resource Metadata (RM WG) разрабатывает стандарт регистров виртуальной обсерватории и спецификацию описания ресурсов. Для этого были рассмотрены несколько индустриальных стандартов, обеспечивающие аналогичные механизмы доступа к данным в телекоммуникационных сетях, и выбран Open Archive Initiative (OAI). Регистры являются для web-сервисов обновляемыми источниками информации об информационных ресурсах. Для описания астрономических ресурсов, а к ним относятся каталоги, цифровые обзоры, базы данных, архивы, программные средства, функционирующие как web-сервисы, в регистрах используются определения Dublin Core (стандарт 3W Консорциума). Описание данных включает (Plane et al., 2004):

- идентификацию: имя и идентификатор ресурса;
- сопровождение ресурса: информация о том, кто поддерживает ресурс (версия, дата релиза) и его наличие;
- описание содержимого: тип данных, область неба, занимаемая данными, спектральный диапазон и т.п.

Метаданные, описывающие ресурс, являются аналогом UDDI (Universal Description, Discovery and Integration) (Clement et al., 2003) для web-сервисов. Программный пакет GLU (Generateur de Liens Uniformes) (Fernique et al., 1998), разработанный в CDS (Centre de Donnees astronomiques de Strasbourg) для разрешения URL-ссылок, впервые использовал набор метаданных для описания астрономических ресурсов. В качестве примера более более подробно остановимся на метаданных, характеризующих общее качество и погрешности измерений физических величин для ресурса, а именно:

- DataQuality – общая оценка ресурса, касающаяся целостности и согласованности, а также уровня документированности, относящегося к погрешностям измерений и калибровок данных. Предлагается следующая градация:

  A – данные прокалиброваны, имеется описание, пригодны для для использования в научных исследованиях (science-ready);

  B – имеются описания и калибровки, однако, пользователь должен проверять данные и, в случае необходимости, проводить повторную калибровку;

  C – некалиброванные данные;

  U – описание ресурса не содержит оценки качества данных.

- ResourceValidationLevel – числовой параметр (от 0 до 4), который характеризует качество описания и соответствие интерфейса доступа к данным стандартам IVOA при использовании его в программных средствах виртуальной обсерватории.

  0 – ресурс не описан, или его описание не соответствует стандарту;

  1,2 – описание и сервис соответствует стандарту; сервис демонстрирует функциональную совместимость со стандартом, и при запросе выдается документ без ошибок;

  3,4 – описание ресурса проверено экспертом, установлены параметры качества, используется программными средствами виртуальной обсерватории.

В отличие от других параметров, описывающих ресурс, значения ResourceValidationLevel устанавливаются не провайдером данных, а администратором регистра. В информационной ин-

фраструктуре, когда запись ресурса может существовать в нескольких регистрах, каждый экземпляр записи может иметь свое значение, в зависимости от стандартов качества, принятых в регистре. Уровни 0, 1, и 2 определены так, что могут назначаться автоматически программным агентом. Для установки значений 3 и 4 требуется участие эксперта.

- ResourceValidatedBy – IVOA идентификатор для регистра, установившего значение для ResourceValidationLevel.
- Uncertainty.Photometric – погрешность фотометрических измерений (в Янских).
- Uncertainty.Spatial – погрешность позиционных измерений (в градусах).
- Uncertainty.Spectral – погрешность определения длин волн (в метрах).
- Uncertainty.Temporal – погрешность шкалы времени.

Наиболее полно описание ресурсов, соответствующее существующей рекомендации, поддерживается регистрами Национальной виртуальной обсерватории США (National Virtual Observatory, NVO), в которых на настоящее время имеется десятки тысяч записей. Регистры NVO позволяют любому пользователю самому выполнять регистрацию ресурса. В Страсбургском центре данных (CDS) действует похожий механизм регистрации и разрешения URL-ссылок на ресурсы – GLU. В системе AstroGrid (UK, e-Science, http://www.astrogrid.org) используют свой регистр с отличающимся набором параметров. Имеются интерфейсы для взаимодействия этих трех типов регистров.

### *2.1.2. Модели данных для астрономии*

Деятельность рабочей группы IVOA Data Models (DM WG) направлена на разработку доменной модели данных, которая в качестве семантического стандарта предназначается для обеспечения интероперабельности программных средств виртуальной обсерватории (Lemson et al., 2003). Интероперабельность в контексте информационных технологий понимается как способность программных систем к сетевому взаимодействию на основе единой модели данных.

В настоящее время астрономическое сообщество в качестве стандарта для хранения и обмена данными использует FITS-формат (Wells et al., 1981). Однако, он не может являться семантическим стандартом, поскольку формат разрабатывался в начале 80-х годов прошлого века и предназначался для обмена данными, поэтому, прежде всего, были стандартизованы параметры, отвечающие за представление информации в разных аппаратно-программных платформах, а ключевые слова, описывающие содержательную астрономическую часть, не были достаточно специфицированы. В стандарте FITS-формата не была определена процедура добавления новых ключевых слов. Это привело к появлению различных вариантов именования ключевых слов и форматов представления их величин, что тем самым усложнило программную интерпретацию параметров наблюдений в заголовках файлов, полученных на разных астрономических инструментах.

Модели данных в информационных технологиях используются для представлений структур данных и их взаимосвязей. Они являются формализацией понятий предметной области и представляет собой набор правил, которые должны применяться при описании данных. Такие правила становятся важными при обмене данными или метаданными в программных приложениях виртуальной обсерватории, и только с их помощью можно гарантировать интероперабельность. Сериализация моделей позволяет провайдерам согласованно описывать предоставляемые данные, а применение классов моделей данных в программных средствах виртуальной обсерватории существенно упрощает их разработку.

Разрабатываются несколько компонентных моделей, являющимися составными частями общей модели, а именно:

- модель данных для наблюдений “Observation Data Model” определяет структуру метаданных, используемых для описания контекста и содержимого файлов, получаемых на аст-

рономических инструментах. В ней описывается область проведения наблюдений, точность регистрируемых величин, обработка, калибровки, условия наблюдений и их планирование (McDowell et al., 2005);

• метаданные пространственно-временных координат (Space-Time Coordinate Metadata или STC) являются моделью данных системы координат (Rots, A., 2007);

• Quantity Data Model – элементарные значения астрономических данных. Астрономические данные состоят из строк и чисел, связанных и организованных определенным образом. Любое числовое значение должно быть соотнесено с физическим понятием, обозначаемым UCD, а также с единицами измерения. В ней определяются ошибки (абсолютные относительные, систематические, случайные), цифровой диапазон значений, размеры пикселя и т.п. (McDowell et al., 2004). Модель данных используется при семантическом описании наборов данных в моделях более высокого уровня;

• Astronomical DataSet Characterisation – характеристики астрономических наборов данных, определяет верхний уровень метаданных, необходимых для описания пространства физических параметров наблюдений или модельных данных, таких как 2D-изображения, кубы данных, списки событий в рентгеновском диапазоне и т.д. Экземпляр модели данных Characterisation может включать описание осей данных, диапазон координат, детали дискретного представления непрерывной величины в наборе данных и разрешение каждой оси (Louys et al., 2007).

• Spectral Data Model описывает структуру спектрофотометрических наборов данных и может использоваться для представления спектров, временных серий, сегментов SED (Spectral Energy Distributions) (McDowell, J., 2007).

### *2.1.3. Семантическое описание понятий*

Определение значений и интерпретация слов, предложений или других языковых форм в астрономическом контексте, а также описание астрономических объектов, типов данных, астрономических понятий и явлений, реализация запросов на естественном языке к астрономическим ресурсам, включая перевод и интернационализацию интерфейсов, относятся к семантике предметной области, которая разрабатывается рабочей группой IVOA Semantics.

В Страсбургском центре звездных данных, где собрана крупнейшая коллекция астрономических каталогов, разработан словарь дескрипторов UCDs, использующийся для установления смысловой связи между обозначениями и астрономическими понятиями. В словарь входит порядка 1500 уникальных описаний содержимого или дескрипторов, отобранных из имеющихся десятков тысяч именований колонок таблиц, хранящихся в CDS. Стандарт для дескрипторов UCD1+ создает согласованный и расширяемый набор из элементарных UCDs (Derriere et al., 2004). Рабочая группа обеспечивает сопровождение словаря UCDs, который используется в протоколах и спецификациях IVOA (Hessman et al., 2007) для установления семантических связей между параметрами запросов и метаданными.

Виртуальная обсерватория при анализе научных данных и поиске новых закономерностей, ориентируется на разработку экспертных систем и баз знаний, которые используют методы искусственного интеллекта, в частности, онтологии. Онтологии также применяются в семантических сетях и технологии программирования и являются формализацией совокупности сведений (данных или программ), отражающих знания экспертов в определенной предметной области. Для знаний характерна внутренняя интерпретируемость, связанность и активность, то есть добавление нового факта может порождать новые связи и правила в базе знаний в отличие от совокупности неких данных. Разработка онтологии для астрономических объектов входит в круг проблем, также решаемых этой рабочей группой IVOA (Cambersy et al., 2007).

*2.1.4. Механизм доступа к данным*

Группа Data Access Layer (DAL WG) разрабатывает стандарты для организации доступа к данным. Клиентские программы, совместимые с этими стандартами, могут использовать web-сервисы, реализующие доступ к данным через инфраструктуру виртуальной обсерватории, а провайдеры данных применять сервисы для публикации данных в регистрах. Стандартами DAL обеспечивается схема, по которой центры данных и обсерватории должны разрабатывать для своих ресурсов сервисы, совместимые со стандартами ВО.

В настоящее время идет работа по созданию спецификации следующих протоколов:

- ConeSearch – поиск данных в каталоге или таблице для заданной области на небесной сфере (Williams et al., 2007);
- Simple Image Access (SIAP) – извлекаются изображения из обзоров или архивов (Tody & Plate, 2004);
- Simple Spectrum Access (SSAP) – спектры и временные серии (Dolensky & Tody, 2004);
- Spectral Line Access (SLAP) – параметры спектральных линий (Salgado et al., 2005);
- Table Access (TAP) – табличные данные (Stebe et al., 2007; Tody, 2007).

Приведенные выше протоколы предназначены для работы с одним источником информации (каталогом, архивом или обзором). В отличие от них протокол доступа к каталогам SkyNodes Interface используется для запросов сразу к нескольким каталогам (IVOA SkyNode Interface, 2004). Он использует расширенное подмножество SQL, называемое Astronomical Data Query Language (ADQL) (Ysuda et al., 2004). ADQL, кроме координатных запросов, поддерживает доступ по протоколам ВО к таблицам, изображениям и спектрам. ADQL, в основном, использует подмножество оператора SELECT с дополнительными функциями, позволяющими определять геометрические типы данных и выполнять операции над ними. В базовом наборе операций ADQL предусмотрены запросы метаданных ресурса о грамматических спецификациях, таблицах, колонках, функциях, изображениях.

## 2.2. Подготовка научных (science-ready) данных

Виртуальная обсерватория использует готовые для научного анализа данные (science-ready), в которых исправлены инструментальные ошибки, с помощью калибровок выполнен переход от инструментальных значений к реальным физическим величинам.

*2.2.1. Качество редукции данных и полнота описания наблюдения*

Как известно, на результаты наблюдений влияет множество факторов: погодные условия (качество изображения, прозрачность), состояние аппаратуры (светоприемников, приборов и телескопа), а также человеческий фактор.

Результаты наблюдений, получаемые на современных цифровых приборах, обычно записываются в файлв в стандартном астрономическом формате. Это - Flexible Image Transport System или FITS (Wells et al., 1981). Он является самодокументируемым форматот, где в файл, кроме наблюдательных данных, записываются еще параметры, характеризующие наблюдение. Чем полнее описаны требуемые идентификации и обработки параметры наблюдения, тем надежнее, в конечном итоге, результат редукции.

При всех своих положительных качествах FITS-формат имеет недостатки. Они связаны с тем, что стандарт был разработан достаточно давно, в начале 80-х прошлого века, и предназначался прежде всего для обмена данными между различными программно-аппаратными платформами, поэтому в нем строго определены только те ключевые слова, которые определяют представление данных.

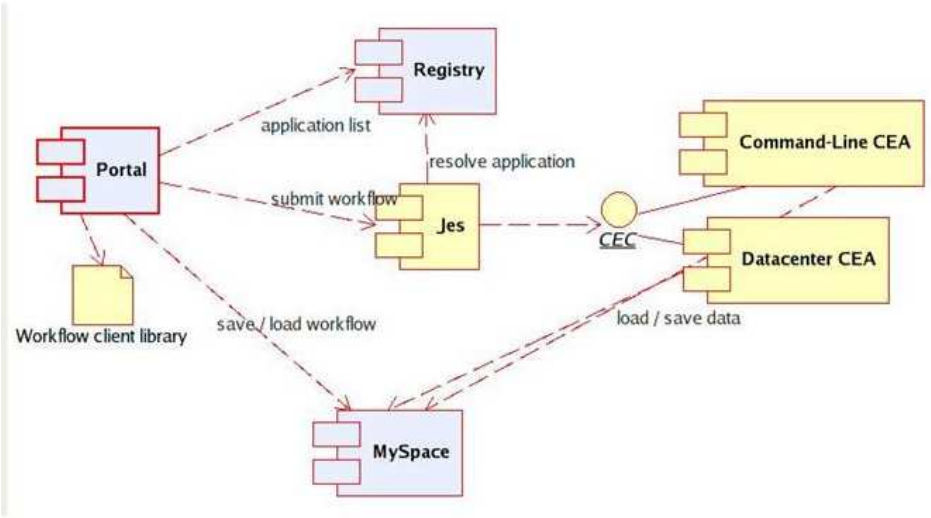


Рис. 1: Схема реализации потока работ в системе AstroGrid.

В начале века были переопределены ключевые слова и расширено описание координатной системы изображений (Greisen & Calabretta, 2002), но для ключевых слов, которые используются для описания других параметров наблюдений, нет четко определенных требований на наличие в описательном заголовке файла. В FITS-стандарте не определены правила генерации новых ключевых слов. За время использования формата в разных обсерваториях появились свои клоны ключевых слов. В результате один и тот же физический параметр наблюдения может по-разному называться в файлах, записанных даже на одном телескопе.

Значения параметров наблюдения попадают в заголовок файла из систем контроля и управления телескопом, прибором, приемником излучения, из интерфейса наблюдателя. Чем выше автоматизация отдельных систем и их взаимодействие, тем меньше наблюдатель вносит вручную значений и, следовательно, в них меньше вероятность пропусков и ошибок.

В ведущих обсерваториях мира достаточно давно используется понятие наблюдательного цикла. Он рассматривается как единый технологический процесс, состоящий из отдельных взаимосвязанных этапов. Сюда входит подача заявок на наблюдательное время, составление расписания, подготовка к наблюдениям, сам процесс наблюдений, архивирование необработанных данных, подготовка и проверка калибровочного материала, верификация правильности и полноты заполнения заголовков файлов, обработка, архивирование научных данных. Такие системы действуют для крупнейших телескопов мира, например, в Европейской южной обсерватории (ESO) специальный отдел Data Management and Operations Division (DMD, http://www.eso.org/org/dmd/) занимается обеспечением наблюдательного цикла. Такие системы требуют достаточных вложений и трудовых затрат, поэтому они не везде используются. К сожалению, еще нет общепринятых стандартов для отдельных частей наблюдательного цикла.

### *2.2.2. Первичная обработка данных*

Обычно обработка наблюдательных данных связана с фиксированной последовательностью операций. Операции с изображениями – это арифметические действия, фильтрация, статистика. Для различных астрономических приборов методы редукции данных отличаются. Последовательность обработки можно формализовать с помощью фиксированных правил.

В IVOA разрабатывается программное обеспечение для распределенных вычислений, пред-

ставляющее собой связанный набор web-сервисов с потоковым принципом выполнения. Поток задач для обработки распределенных данных состоит выполнении процедуры, заданной пользователем, с использованием программных сервисов, связанных вместе циклами, условными переходами и т.п. для более сложных операций. Для компонентов потока задач определены правила вызова сервиса, структура входных и выходных данных.

Программная система AstroGrid (http://www.astrogrid.org), реализованная на основе стандартов IVOA, состоит из нескольких компонентов, посредством которых реализуются вычисления и доступ к данным в Интернете. Разработанный в этой системе механизм управления потоком задач может послужить прототипом для динамической обработки наблюдений в архивах, включенных в инфраструктуру виртуальной обсерватории (см. рис.1) (Walton et al., 2006). Для этого необходима стандартизация последовательности шагов обработки (pipeline) для разных инструментов, создания библиотек таких процедур, а также разработка приемов автоматического контроля качества калибровочных данных, используемых в редукции.

#### *2.2.3. Хранение и сохранность данных*

Астрономические явления часто носят переменный характер на разных временных интервалах, поэтому долговременное хранение наблюдений обычно входит в компетенцию обсерваторий. Появление цифровых приемников излучения обеспечило астрономов большим количеством данных, а цифровые записывающие устройства - компактными средствами хранения. Однако информацию уже невозможно рассматривать человеческим глазом, как это можно было делать с фотографическими пластинками.

Цифровые носители требуют специального оборудования для раскодирования их содержимого. При считывании, записи и хранении данных могут возникнуть ошибки, потеря информации. Конечно, надежность устройств растет, а ошибки, возникающие при чтении и дублировании носителей информации, можно контролировать программно. Но компьютерное оборудование меняется с такой скоростью, что время физического разрушения носители информации оказывается больше, чем время жизни устройства считывания. Это требует постоянного отслеживания состояния систем хранения и переписывания данных на новые носители. Поэтому для сохранности данных в архивах необходимо пересматривать технологию хранения раз в несколько лет и переносить содержимое с устаревших носителей на более новые по технологии. Такую операцию приходится производить раз в 3-5 лет.

Примерно так выглядит круг проблем, которые необходимо решать при включении архива наблюдений в виртуальную обсерваторию.

## 3. Реинжениринг наблюдательных данных

В Специальной астрофизической обсерватории более 20 лет ведется цифровой архив наблюдательных данных. При включении астрономических архивов в виртуальную обсерваторию нужно учитывать определенные требования к уровню подготовки данных (science-ready), а также обеспечивать web-сервисы, совместимые со стандартами IVOA, для работы с ними. В нашем случае, когда архив данных состоит из разнородных цифровых коллекций, накапливаемых в течение почти двух десятков лет, доведение данных до требуемого уровня стандартизации форматов и полноты описаний параметров требует достаточно объемной работы.

Если в инженерные задачи входит преобразование продукта инженерии из-за изменения требований или стандартов, когда в соответствии с этим выполняется проектирование, разработка, тестирование и реализация новой версии, то такой комплекс работ называется реинжинирингом. Унификацию данных архива до уровня совместимого со стандартами, используемыми астрономическим сообществом, будь то FITS-формат или новые спецификации IVOA, можно также назвать реинжинирингом.

Таблица 1: Общий архив наблюдательных данных САО РАН (CD/DVD-диски)

| Архивы | Число CD/DVD-дисков | Среднесуточн. поток данных | Объем архива | Число записей |
|---|---|---|---|---|
| Оптика | 150+150(копий) | $\sim$ 150 MB | 309 GB | $\sim$ 190000 |
| Радио | 7 | $\sim$ 4 MB | 4 GB | $\sim$ 46000 |
| Личные архивы | 30 копий | | | |

### 3.1. Общий архив наблюдений САО РАН

Для исследования небесных объектов на телескопах обсерватории используются разные наблюдательные приборы, каждый из которых связан с определенным компьютерно-аппаратным комплексом или, говоря иначе, системой сбора. Для разных систем сбора форматы цифровых данных имеют свой набор параметров для описания наблюдений и отличаются друг от друга. Локальный архив – это цифровая коллекция данных, получаемых одной системой сбора. В настоящее время в общем архиве обсерватории их 16. В таблице 1 приведена информация об объеме данных архива, количестве записей и дисков.

В архиве хранятся наблюдательные данные, полученные с помощью наблюдательных методов, использующихся и использовавшихся на телескопах обсерватории и дополнительная информация. Наблюдательные данные – файлы с наблюдениями небесных объектов, сервисные файлы, используемые для коррекции инструментальных ошибок и калибровок, журналы наблюдений, сопутствующая информация, подготовленная наблюдателем. Дополнительная информация – текстовые справочные файлы, информация, используемая для идентификации диска, программное обеспечение, связанное с содержимым диска, контрольные суммы.

С 1994 года цифровые данные, получаемые на инструментах обсерватории, хранятся на оптических дисках. На диски записываются наблюдательные данные и дополнительная информация, подготовленная наблюдателями и администратором архива. На рисунке 2 представлены объем и темпы прироста цифровых наблюдательных данных, накапливаемых в архивной системе.

### 3.2. Информационно-поисковая система общего архива наблюдений

Работа по стандартизации архивных данных в обсерватории ведется с конца 80-тых, когда впервые был разработан FITS-заголовок для данных ПЗС-камеры (Витковский и др., 1988). Затем FITS-формат стал использоваться для описания и хранения наблюдательных данных, полученных на других приборах.

В конце 90-х отдел информатики приступил к созданию архивной системы для накопления и постоянного хранения цифровых данных, получаемых на телескопах обсерватории, а также поиска и доступа к ним из Интернета. Для нее была предложена архитектура, состоящая из трех взаимосвязанных уровней:

- накопление – каскадная схема архивизации, где поток данных от любой системы сбора направлялся на общий файл-сервер, а затем запись на носители для постоянного хранения;
- хранение – библиотека оптических дисков и хранилище на жестком диске выделенного архивного сервера. Первоначально для хранилища предполагалось использовать робот оптических дисков, но потом из-за технических сложностей, связанных с поддержкой этого устройства в разных версиях LINUX, а также с ростом объема и быстрым понижением стоимости носителей на жестких дисках от этой идеи отказались;

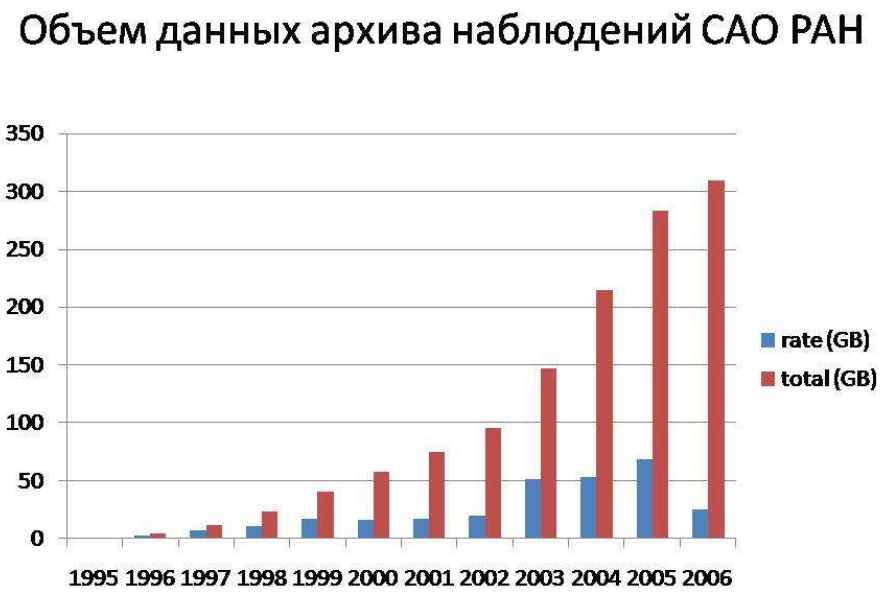


Рис. 2: Общий объем и ежегодный прирост данных в архиве наблюдений. На рисунке представлены данные об архивизированных файлах в информационно-поисковой системе архива.

• поиск и доступ – информационно-поисковая система (ИПС) на основе системы управления реляционными базами данных с пользовательским web-интерфейсом.

*3.2.1. Прототип информационно-поисковой системы*

В отделе информатики совместно с коллегами из ЮгИнфо РГУ приступили к разработке прототипа ИПС для поиска наблюдений в архиве и доступа к данным из Интернета (Vitkovskij et al., 2000). Архив представлял собой коллекцию CD-дисков с разнородной цифровой информацией, где наблюдательные данные были записаны в пяти форматах, включающих FITS-формат, внутренний формат системы MIDAS, двоичные файлы без описательной информации, двоичные файлы с текстовым описанием заархивированные в tar-архивы, радиоастрономические наблюдения в формате FLEX (Верходанов и др., 1993).

При разработке информационно-поисковой системы общего архива наблюдений обсерватории мы руководствовались следующими правилами:

• в архиве не меняются форматы хранимых данных; в каком формате они поступили на вход архива, в таком они выдаются при запросе;

• в архиве хранятся необработанные данные, записанные на оптических дисках;

• логической единицей хранения в архиве является наблюдение;

• данные имеют двухлетний период исключительного авторского права заявителей наблюдательной программы на основании Положения об архиве наблюдательных данных САО РАН, затем они открываются для свободного копирования.

Разработка схемы таблиц информационно-поисковой системы (ИПС) проводилась в ЮгИнфо РГУ. Архив имел двухуровневую организацию. ИПС образовывала первый уровень, который

обеспечивал запросы по параметрам наблюдений, а второй уровень являлся хранилищем файлов. Для каждого файла в ИПС из параметров наблюдений формировался сервисный информационный блок, который включал также информацию для идентификации файла. Такая архитектура архивной системы позволила снять ограничение на форматы наблюдательных данных. В прототипе рассматривались данные, записанные на оптические диски в FITS-формате. Информационно-поисковая система базировалась на системе управления базами данных Oracle. Сервер базы данных располагался в Ростове. Запросы и копирование данных выполнялись с помощью web-интерфейса. Тестирование ИПС проводилось на радиоастрономических данных, полученных на РАТАН-600.

В сервисный информационный блок было заложено избыточное количество параметров (весь заголовок FLEX-формата), поэтому возникли сложности с наполнением таблиц ИПС, поскольку значительная часть параметров не обеспечивалась в других локальных архивах из-за специфики методов наблюдений, а также из-за пропусков и ошибок в значениях параметров наблюдения при регистрации. В системах сбора обычно не предусматривается проверка правильности заполнения описательной части наблюдательного файла. Другая проблема, в связи с которой прототип ИПС не был применен ко всему архиву, связана с выполнением Положения об архиве наблюдений обсерватории, а именно, с авторским правом заявителей программ на полученные данные.

### *3.2.2. Поиск данных по дате наблюдения*

Идеи, заложенные в реализацию прототипа поисковой системы архива, послужили основой для разработки и создания существующей архивной системы. Прежде всего мы отказались от того, чтобы в таблицы ИПС заносились все параметры заголовков файлов. Чтобы обеспечить доступ ко всему архиву, требовалось выделить те параметры наблюдательных файлов, которые присутствовали бы во всех локальных архивах, а также были наиболее устойчивы к пропускам и ошибкам со стороны наблюдателей.

Таким параметром оказалась дата наблюдений, поскольку наблюдатели обычно сохраняют получаемые в течение одной ночи файлы в каталоге, название которого содержит дату. Такая структура каталогов автоматически переносится на общий файл-сервер, куда копируются данные сета наблюдений. Затем из сетов наблюдений, полученных на одном приборе, формируется образ архивного диска, который по мере заполнения записывается на CD/DVD-носитель. Из этого естественным образом сложились простые правила, которые определяют внутреннюю структуру архивного CD/DVD-диска:

- метка тома (файл нулевой длины);
- каталоги с наблюдениями (иногда имеются вложенные каталоги);
- в одном каталоге записываются данные одной наблюдательной ночи;
- название каталога должно содержать дату наблюдения
- семантическая единица архива - один файл с наблюдательными данными.

Эти условия не ограничивают форматы файлов, поэтому данные нового прибора, установленного на телескопе, легко добавляются в архив и информационно-поисковую систему, что обеспечивает масштабируемость архивной системы в смысле форматов файлов (так называется возможность добавления новых функций или данных без изменения структуры самой программной системы). При выполнении перечисленных выше правил можно добавлять в ИПС наблюдения, полученные в других обсерваториях, но запрос к данным будет ограничиваться датой и методом наблюдения.

Архивная система включает в себя информационно-поисковую систему на основе реляционной СУБД (системы управления базами данных) и хранилище данных, состоящее из коллекции оптических дисков и области хранения на диске выделенного сервера. Для обеспечения

сохранности данных каждый архивный диск представлен в коллекции в двух экземплярах, кроме того, еще имеется копия данных на жестком диске, тем самым при возникновении ошибок чтения на одном из типов носителя можно восстановить информацию с другого носителя.

Область хранения наблюдательных данных на жестком диске архивного сервера имеет физическую и логическую организацию. Каждый диск помещается в каталог с названием диска, соответствующим его архивному номеру – CDxxx. Ниже уровнем находятся каталоги, именованные по дате наблюдений и содержащие наблюдения на соответствующую дату. На физическую структуру архива накладывается логическая, в которой отображается распределение дисков по локальным архивам (методам наблюдений).

Вышеперечисленные правила определяют процедуру верификации CD-диска перед помещением его в ИПС, а именно, проверяется:

- сколько локальных архивов записано на диске;
- хранятся ли данные одной ночи в одном каталоге, именованном по дате наблюдений;
- соблюдается ли в каталоге правило записи наблюдения в один файл.

Помещение диска в архивную систему начинается с копирования в буфер, где проверяется соответствует ли его содержимое принятой структуре, а затем при записи в область хранения на винчестере выполняется верификация. При записи диска в хранилище выполняются следующие проверки и преобразования:

- если на диске записано несколько локальных архивов, то обработка диска проводится в несколько проходов - столько, сколько записано методов на диске. Символьная ссылка на один и тот же диск устанавливается тогда в нескольких соответствующих каталогах логической структуры области хранения;
- данные одной ночи переписываются в один каталог, если на копируемом диске это не соблюдается;
- если имя каталога не содержит дату наблюдений, то он переименовывается так, чтобы дата присутствовала в названии;
- наблюдения извлекаются из tar-архивов, если таковые имеются;
- файлы заново компрессируются, если метод компрессии отличается от bzip2;
- выполняется преобразование файлов во внутреннем формате системы MIDAS (файлы с расширением bdf) в FITS-формат;
- производится анализ FITS-заголовков, если в заголовках указан другой инструмент, чем тот, к которому относится диск, то принадлежность файла локальному архиву изменяется.

Верификация выполняется программно с сохранением сообщений об ошибках и действий с дисками в протокольных файлах. Дополнительные операции с каждым диском фиксируются в программе на языке bash (командная оболочка системы UNIX), который используется при полном или частичном восстановлении хранилища и ИПС.

Все эти проверки выполняются программами-фильтрами перед созданием текстовых файлов, отражающих структуру диска. Эти текстовые файла получаются при выполнении команды ls системы UNIX. Они являются частью архивной системы, сохраняются и используются при наполнении таблиц базы данных.

Поскольку наиболее надежным параметром для идентификации файла является дата наблюдений, определяемая из имени каталога с данными одной ночи, то именно по дате, начинается наполнение таблиц ИПС.

Первый вариант ИПС обеспечивал копирование по HTTP-протоколу данных, включенных в информационно-поисковую систему, просмотр заголовков файлов с параметрами наблюдений и содержимого файлов. Запросы выполнялись по дате наблюдений и локальному архиву/методу наблюдений. Схема таблиц этого варианта ИПС представлена на рисунке 3.

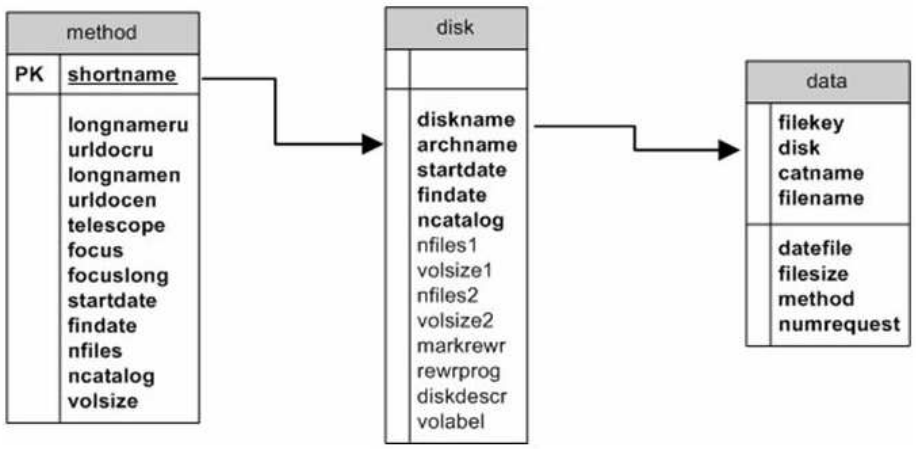


Рис. 3: Схема таблиц версии базы данных архива, обеспечивающей доступ к данным по дате наблюдений

Из расписаний наблюдений, опубликованных на сайте обсерватории в разделе "Телескопы", по ссылкам, обозначенным "Данные", можно отправить запрос к ИПС на интервал дат выбранного сета.

Для архива наблюдений используется специализированный сервер, на котором размещены ИПС на основе СУБД (Oracle 9i trail-версия) и хранилище файлов. Интерфейс пользователя и on-line доступ к данным выполнен на основе спецификаций CGI-интерфейса (Common Gateway Interface), DBD (DataBase Driver) и DBI (DataBase Interface) интерфейсов к СУБД. ИПС имеет две схемы – основную и тестовую (для разработки и тестирования). Поддержка двух схем базы данных, дублирующих в какой-то степени одна другую, обеспечивают дополнительную устойчивость системы к ошибкам чтения данных. Таблицы ИПС пополняются по мере поступления CD-дисков с новыми данными.

### *3.2.3. Обеспечение наиболее часто используемых запросов к архиву наблюдений*

По опросам пользователей мы определили наиболее часто используемые типы запросов для выборки архивных данных, а именно, - по дате наблюдения, прибору, типам файлов, координатам наблюдаемого поля/объекта, имени астрономического объекта, программе наблюдений, заявителю программы и наблюдателям, принимавшим участие в наблюдениях.

Как видно из анализа данных, помещенных в архив, для описания наблюдений на разных приборах БТА применяются наборы параметров, имеющие общую часть, которая включает информацию об объекте, программе и т.п., и технические характеристики, присущие конкретному прибору. В описание входит до 75 параметров (на радиотелескопе РАТАН-600 - до 289), они сохраняются в заголовке файла. Значения этих параметров формируются в системах управления телескопом, инструментом, а также в системе сбора данных. Часть параметров поступает в заголовок файла автоматически, часть заносится наблюдателем. Запрос по дате наблюдения реализуется ко всему архиву, но другие типы запросов можно выполнить только к части локальных архивов из-за отсутствия необходимых параметров в заголовках наблюдательных файлов.

Сложность заполнения таблиц ИПС необходимыми для стандартных запросов параметрами состоит в том, что:

- при модернизации систем сбора и приборов меняются форматы данных и, как прави-

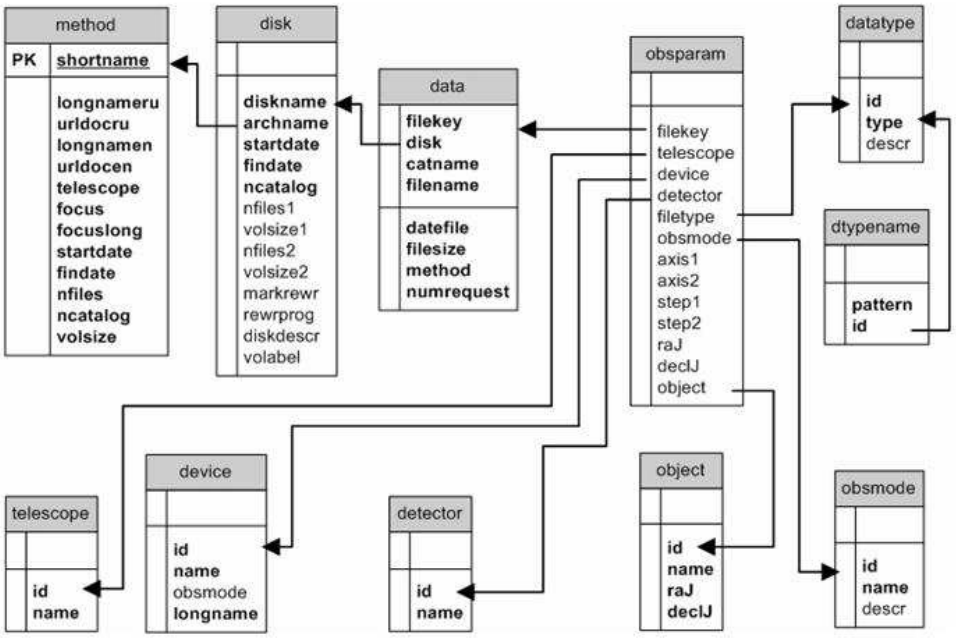


Рис. 4: Схема таблиц версии поисковой системы общего архива наблюдений САО РАН для стандартных запросов

ло, в локальном архиве имеется несколько версий формата, которые отличаются по набору ключевых слов, а также форме записи их величин;

- разные системы сбора формируют заголовок файла с отличающимися по названию ключевыми словами, но обозначающими одну и ту же физическую величину. К примеру, дату наблюдения в разных цифровых коллекциях можно получить из значений следующих ключевых слов: “DATE”, “DATE-OBS”, “Date of observation”, “OBS-DATE”.

По этим причинам программный фильтр для синтаксического разбора, анализа и извлечения значений параметров из заголовков файлов реализован с использованием дополнительной таблицы в схеме базы данных, которая связывает названия ключевых слов и параметры наблюдения. Такая таблица называется словарем или тезаурусом и содержит все ключевые слова локальных архивов, обнаруженные при разборе FITS-заголовков файлов, их семантические значения, определяющие физический смысл величины, а также связь величины и параметра ИПС. На рисунке 4 приведена схема таблиц для реализации стандартных запросов.

В заголовках достаточно большой части файлов архива, особенно той, которая относится к первой половине 90-х годов, отсутствуют параметры для стандартных запросов. Мы старались в таких случаях, где возможно, использовать дополнительную информацию, которая имеется в полном имени файла, в которое включены каталоги от корневого каталога оптического диска. По нему определялась дата наблюдения, метод компрессии, тип изображения (темновой кадр, байес, плоское поле, объект), тип фильтра при широкополосной фотометрии, формат записи данных и т.п. Значениям параметрам ИПС, извлеченным из имени файла, придавался самый высокий приоритет по сравнению с информацией из заголовка, то есть в конфликтной ситуации, когда имеются две величины для параметра, в таблицу заносится значение из имени файла.

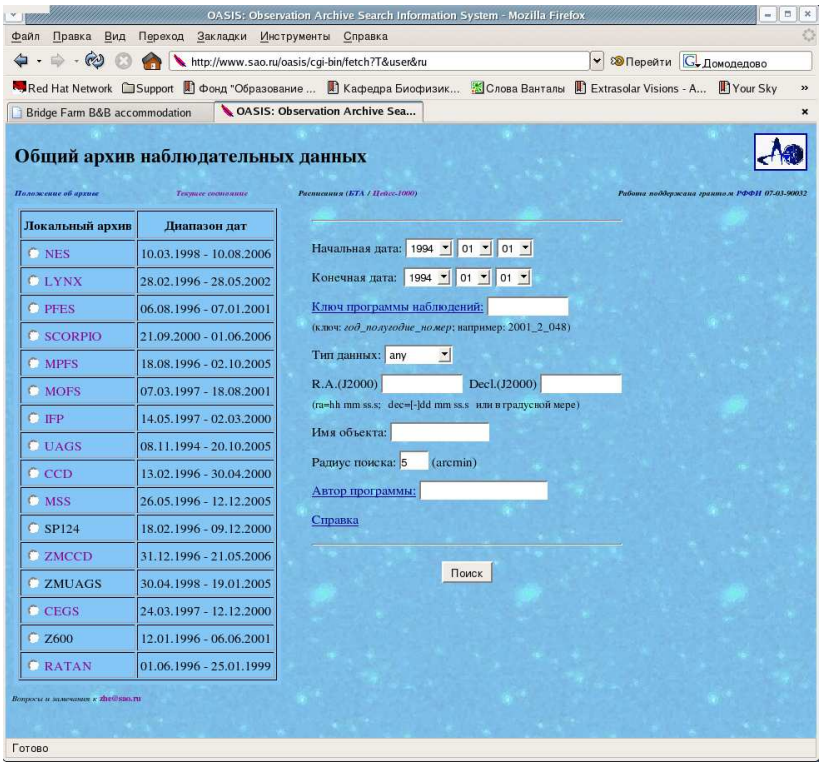


Рис. 5: Web-интерфейс общего архива наблюдательных данных

Дополнительная информация о наблюдениях имеется в расписаниях и журнальных файлах. Она пока не используется при добавлениях в таблицы параметров, которые пропущены в наблюдательных файлах.

### *3.2.4. Web-интерфейс общего архива наблюдений*

Вид пользовательского web-интерфейса общего архива наблюдений обсерватории представлен на рисунке 5, также есть англоязычный вариант интерфейса. Результаты запроса для отмеченного локального архива выводятся в новое окно браузера. Интерфейс администратора архива отличается выводом дополнительной информации для локализации файлов в области хранения. Интерфейс пользователя генерируется скриптовой программой на Perl и динамически отображает диапазон дат помещенных в архив наблюдений и включенные в поисковую систему локальные архивы. С его помощью реализуются стандартные запросы к архивным данным.

Если в html-форме при запросе указаны только даты, то выбираются все наблюдения, а также журнальные и вспомогательные файлы, попавшие в диапазон дат. При задании координат поиск FITS-файлов производится в квадрате со стороной, равной заданному удвоенному радиусу поиска. Центр области задается введенными координатами (прямое восхождение и склонение на эпоху 2000.0). В результате запроса выдаются файлы, у которых координаты, заданные в параметрах FITS заголовка и приведенные на 2000.0, попадают в указанную область. Если задается только координата прямого восхождения, то поиск файлов выполняется в полосе склонений [-90° – +90°], а если вводится только склонение, то поиск файлов выполняется в полосе прямых восхождений [0° – 360°].

При запросе наблюдательных данных по имени объекта его координаты извлекаются с помощью web-сервиса Sesame (http://cdsws.u-strasbg.fr/axis/services/Sesame) из базы данных Simbad, при этом, если имя не обнаруживается, то выполняется поиск в Vizier, а затем в

NED. Следует отметить, что если заданы координаты и имя объекта, то поиск в архиве будет выполняться по координатам, имя объекта в этом случае игнорируется.

При поиске наблюдений по фамилии наблюдателя или заявителя программы допускается неполный ввод фамилии. Если при этом возникает неоднозначность, то поиск в архиве осуществляется по первой фамилии, совпавшей с введенным шаблоном. Например, введено “mon”, этому шаблону соответствуют: “Monin”, “Montmerle”. Поиск производится по “Monin”.

Поиск наблюдений, полученных по какой-либо заявленной программе, производится по ключу программы, который представляет собой строку, включающую год, полугодие и порядковый номер программы в полугодии. Для справки в отдельное окно браузера выводится список программ с соответствующими ключами.

## 4. Анализ параметров, используемых в стандартных запросах

Если параметры, по которым производится обращение к поисковой системе, отсутствуют в заголовке файла, то он не войдет в результат запроса. Такая ситуация не является редкой для архивных данных, особенно для старых файлов. Пропущенную, потерянную при перезаписях информацию, а также ошибки, вносимые в данные наблюдателем, можно в ряде случаев дополнить или исправить, используя другие источники, описывающие наблюдения. Рассмотрим правила, которыми мы руководствовались при наполнении таблиц поисковой системы.

• Дата наблюдений в архивных файлах оказалась наиболее устойчивой к пропускам и ошибкам со стороны наблюдателей. Мы получаем ее из названия каталога с наблюдениями ночи, который входит в полное имя каждого файла, анализируемое программно при помещении в архив. Дата наблюдений из заголовков файлов является менее надежной, поскольку отмечались пропуски в значении ключевого слова, определяющего дату, а также при переходе на новый формат записи даты в течение 2000 года некоторыми системами сбора записывались ошибочные значения. Значения даты наблюдений, которые имеются в таблицах информационно-поисковой системе можно использовать для исправления ошибок в заголовках файлов. Ко всем архивным файлам можно обратиться по дате наблюдения.

• По расширению имен файлы на архивных дисках мы разделяем на наблюдательные, журнальные, вспомогательные (по содержимому можно заключить, что файлы относятся к наблюдениям) и без категории (по содержимому не можем отнести к наблюдательной информации). Отнесение файла к одной из этих категорий производится с помощью справочной таблицы, в которой собраны все встречающиеся в архиве расширения имен файлов. Так на текущий момент расширения .bdf, .mt, .fts, .fits, .tar приписаны наблюдательным файлам, .tbl, .log, .plog, .base, .pro, .dbf, .lst имеют журнальные файлы, а остальное - вспомогательные и без категории. На текущий момент к файлам без категории отнесены и наблюдения РАТАН-600, поскольку для них пока не производится разбор заголовка. Журнальные, вспомогательные файлы и файлы без категории можно извлечь из архива только по дате наблюдения, наблюдательные файлы - по стандартным запросам. Поскольку встречаются журнальные файлы с расширением .mt, то дополнительно анализируются заголовки FITS-файлов для того, чтобы разделить таблицы и изображения. Табличные файлы помечаются затем, как журнальные.
Всего файлов - 235578, наблюдательные данные, включая и радиодиапазон, составляют $\sim$91%, журналы наблюдений $\sim$1%, файлы без категории $\sim$1.5%.

• К наблюдательным файлам относятся научные данные с наблюдениями изучаемых небесных объектов и сервисные данные, используемые при редукции (байесы, темновые кадры, плоские поля, стандарты).
Разделение наблюдательных данных на типы выполняется при анализе имен файлов и ключевых слов в заголовках. При определении типа изображения производится проверка на наличие в значении ключевого слова сочетаний символов, которые совпадают с сочетаниями, извле-

ченными из справочной таблицы. Если такое совпадение имеется, то файлу устанавливается признак соответствующего типа наблюдательных данных. Такой же поиск выполняется и для имени файла.
Если имеется информация о типе данных, как в имени файла, так и в ключевых словах, то при программном анализе приоритет отдается имени файла.
Из 169890 наблюдательных файлов, полученных на оптических телескопах, 96% записаны в FITS-формате. Из этих данных:

- файлы с наблюдениями объектов (OBJ) составляют ∼55%,
- байесы (BS) ∼21%,
- темновые кадры (DK) ∼3%,
- плоские поля (FF) ∼10%,
- стандартные лампы (ST) ∼10%,
- а также те файлы, для которых тип данных не определяется программным алгоритмом (undf) ∼1%.

• Приборы, используемые на телескопе, формируют разные по структуре данные. Так прямой снимок включает информацию (координаты и интенсивности в широкой спектральной полосе) о многих небесных объектах, попавших в наблюдаемую область, а спектрофотометрические данные относится к одному объекту, если не говорить о мультиобъектной спектроскопии. Для реализации запроса по координатам мы разделили наблюдательные файлы по типу наблюдений. В специальной справочной таблице поисковой системы устанавливается отношение прибор - тип наблюдения. Для приборов с несколькими режимами наблюдений (например, SCORPIO) тип наблюдения устанавливается по ключевому слову в заголовке и таблице, содержащей все встречаемые значения ключевого слова и связь их с типом наблюдений. Разделение по типу наблюдений в архиве выглядит так:

- прямые снимки ∼55%,
- эшелле-спектры ∼10%,
- спектры, полученные с длинной щелью, ∼17%,
- наблюдения интерферометром Фабри-Перо ∼12%,
- мультиобъектная спектроскопия ∼15%
- и файлы, для которых тип не определяется программно, – $<0.1\%$.

• Запрос по координатам выполняется к файлам, содержащим наблюдения небесных объектов. В этом типе запроса участвуют координаты, извлеченные из заголовка файла и приведенные к стандартной эпохе 2000.0. Для спектров - это координаты объекта, а для прямых снимков - центра области. В последнем случае при выборке данных надо учитывать размер поля, который вычисляется по количеству и угловому размеру пикселей матрицы ПЗС-камеры. Эти значения извлекаются из соответствующих ключевых слов заголовка, но оказывается, что для ∼55% прямых снимков в архиве угловой размер пикселя равен некоторому фиктивному значению - единице. По этой причине мы решили пока не учитывать для прямых снимков размер области и ограничиться координатами центра.

• Для получения координат по имени объекта используются web-сервисы CDS и NED. Если значения координат в ключевых словах заголовка файла отсутствуют, то можно попытаться получить координаты по имени объекта.

• Имена наблюдателей и авторов программ извлекаются из ключевых слов OBSERVER и AUTHOR. Как правило, эта информация заносится в заголовок наблюдателями, при этом используются сокращения фамилий. Чтобы разобраться с многочисленными синонимами, нам пришлось извлечь из FITS-заголовков наблюдательных файлов список всех возможных вариантов записей фамилий и сокращенных обозначений участников наблюдений.

В файл попало 838 уникальных сочетаний. Некоторые фамилии имеют до десятка вариантов записи фамилии. Каждой персоне, оказавшейся в списке, присвоен номер. Этот же номер присваивался и соответствующему варианту записи. Всего оказалось 206 персон, из которых только заявителями программ является 154, наблюдателями – 133, авторы и наблюдатели – 81. Список сочетаний и список персон послужили основой для справочной таблицы. Выборка данных по автору программы или наблюдателю выполняется посредством этой таблицы, то есть, по совпадению фамилии определяется идентификатор, а затем уже идет поиск с ним по наблюдательным данным.

• Сложность организации запроса по названию наблюдательной программы состоит в том, что названия в заголовки файлов заносятся наблюдателями в произвольной манере, так, например, это может быть смысловое сочетание (“Исследование GRB”), общепринятое сокращение (“GRB”) или неполное название, но не то, которое фигурирует в расписании наблюдений. В таких случаях одной наблюдательной программе в архиве может соответствовать несколько названий, которые, однако, никак не совпадают с названием в расписании наблюдений. В списке программ (∼920 названий), которые были извлечены из наблюдательных файлов (ключевое слово PROGRAM в заголовке), не оказалось ни одного совпадения с названиями программ из архива расписаний наблюдений на БТА (со второго полугодия 1992 по первое полугодие 2007). Общепринятой практикой в других обсерваториях является присвоение программе наблюдений уникального идентификатора. Из архива расписаний составлена таблица (2515 названий). Каждой программе приписан символьный ключ, включающий год, полугодие и порядковой номер программы в полугодии. Выборка данных, относящихся к сету наблюдений, производится по ключу программы.

Нужно учитывать, что стандартные запросы, кроме выборки данных по дате наблюдения, реализуются к части наблюдательных файлов из-за отсутствия параметров в заголовках файлов. Приведем статистику наличия параметров, которые используются в запросах. Отметим, что в ней не учитываются ошибочные значения параметров, а только пропущенные:

- 17% - нет значений координат в ключевых словах заголовка,
- 4% - нет имени объекта,
- 30% - не определено название программы,
- 29% - не указан заявитель программы,
- 29% - не указаны наблюдатели, участвовавшие в наблюдениях.

## 5. Схема таблиц базы данных архива

Система управления базами данных PostgreSQL является свободно распространяемым некоммерческим программным продуктом, поддерживает стандарт ISO и ANSI языка запросов SQL-92 (Грабер, 2003), обеспечивает транзакции, поддерживает представления и сложные структуры данных, создаваемые пользователями, то есть ничем не уступает коммерческим системам (Бартунов, 2005). Эта СУБД используется в программных системах и приложениях виртуальной обсерватории, таких как AstroGrid (Walton et al., 2007), TOPCAT (Taylor, 2007). Для нее разработаны встроенные пакеты процедур, обеспечивающие работы с астрономическими координатами, поддерживающие пикселизацию неба. По этим причинам решено поисковую систему архива перевести в СУБД PostgreSQL. Исходя из накопленного опыта эксплуатации поисковой системы, решено внести изменения в схему таблиц. Она была дополнена представлениями и ограничениями целостности для обеспечения непротиворечивости данных.

Приведем детальное описание схемы таблиц базы данных поисковой системы (см. рис. 6), развернутой в СУБД PostgreSQL. По способу занесения информации таблицы условно можно разбить на три группы по частоте добавления в них новых записей.

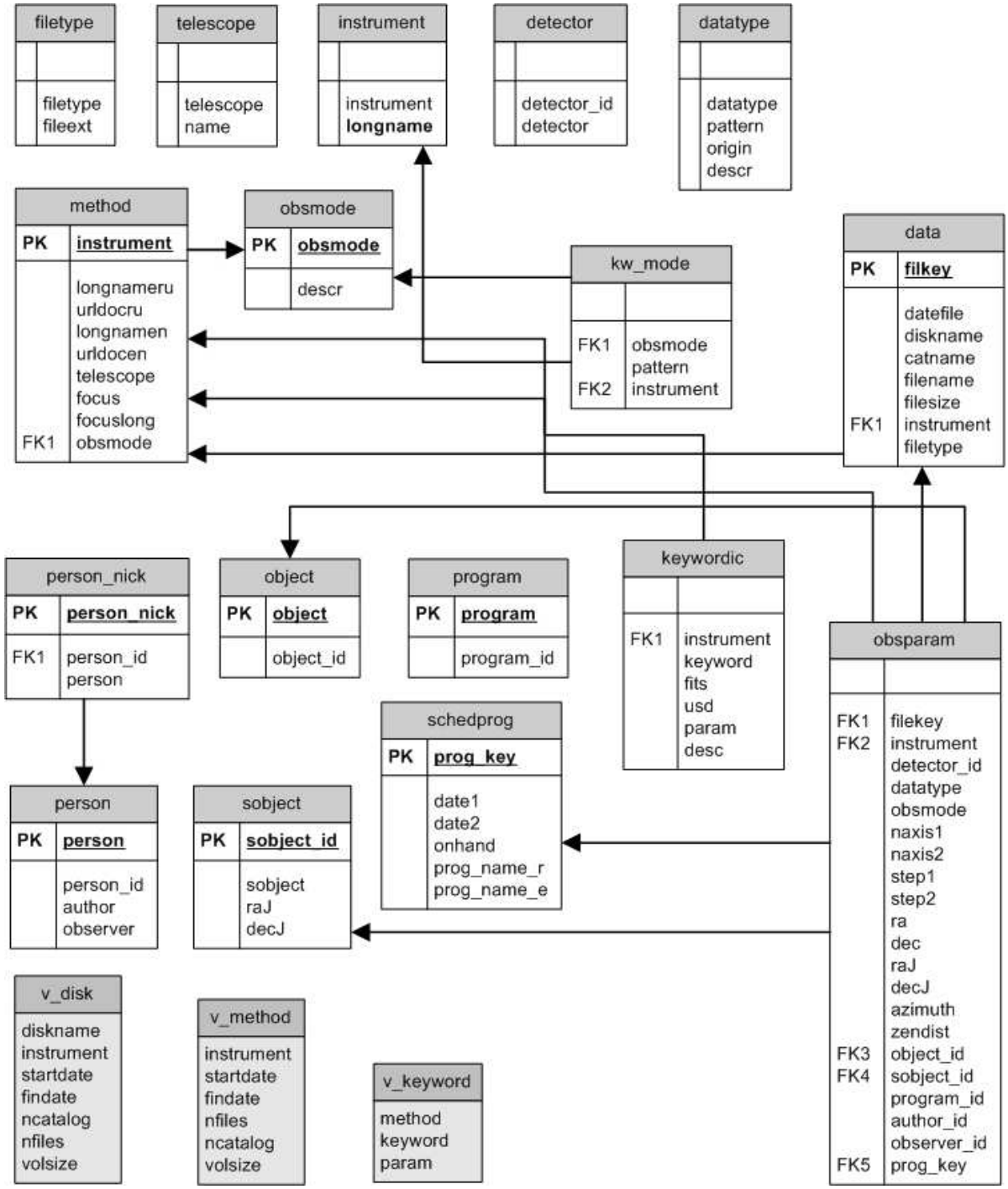


Рис. 6: Схема таблиц и представлений информационно-поисковой системы общего архива наблюдений обсерватории

Таблица 2: Структура записи таблицы filetype. Определение типа архивного файла

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| filetype.filext | расширение имени файла | NOT NULL UNIQUE |
| filetype.filetype | тип данных (obs\|log\|aux) | — |

К первой группе относятся справочные таблицы, информация в которых не меняется. Они заполняются при создании схемы базы данных. Новые записи могут появиться только, когда в информационно-поисковую систему добавляется новый локальный архив. Такие добавления происходят не часто, поэтому эти таблицы можно считать статическими.

Ко второй группе относятся справочные таблицы, которые могут пополняться новыми записями при анализе впервые вносимого в поисковую систему CD/DVD-диска, например, новая программа наблюдений, наблюдатель, тип файла и т.п. Добавление записи происходит программно. Информация, которая заносится в справочные таблицы, получена в результате визуального и программного анализа архивных дисков и заголовков файлов.

Третья группа – это таблицы с данными о каждом архивированном файле. Записи добавляются в них при внесении нового CD/DVD-диска, и при заполнении полей этих записей используются таблицы первых двух групп.

Особое место занимает таблица, связывающая атрибуты таблиц, хранящих информацию о наблюдательных файлах, с ключевыми словами FITS-заголовков и UCD.

В схеме ИПС есть представления, которые формируются динамически средствами системы управления базами данных на основе справочных таблиц.

В приложении А приведен список программ, обеспечивающих наполнение таблиц поисковой системы архива наблюдений и реализующих запросы.

### 5.1. Статические таблицы поисковой системы архива

В базе данных поисковой системы архива несколько статических справочных таблиц. Информация, которая в них собрана, предназначена для анализа и интерпретации значений ключевых слов в FITS-заголовках и имен файлов.

Таблица filetype используется при анализе списка имен файлов архивного CD/DVD-диска, сформированного командой ls, для разделения данных по расширению имени файла на наблюдательные, журнальные или вспомогательные, помечаемые затем в таблице data (см. табл.19) как “obs”, “log”, “aux” соответственно. Структура записи filetype представлена в таблице 2. Содержит 33 записи.

Отображаемая web-интерфесом информация о локальных архивах собрана в таблице method (см. табл. 3). Для поля (атрибута) записи telescope установлена ссылка на соответствующий атрибут таблицы telescope (см. табл. 4) с ограничинием “ON UPDATE CASCADE” для поддержки непротиворечивости данных при обновлении родительской таблицы (Грабер, 1996). При внесении изменений в таблицу telescope будут обновлятся в таблице method названия телескопов. Аттрибут obsmode ссылается на таблицу obsmode (см. табл.8). В нем определяются режимы наблюдений приборов.

В архиве хранятся наблюдения, полученные на телескопах САО РАН. В заголовках наблюдательных файлов встречается разные варианты записи названий телескопов, например: “BTA”, “BTA 6-m”, “6m”, “6-m”. Они собраны в одну таблицу для установления соответствия принятому в поисковой системе названию для данного телескопа. В таблице 13 записей.

Название телескопа в зависимости от системы сбора записываются в разные ключевые слова заголовка файла. Связь “название телескопа” – “ключевое слово” – “система сбора” сохраняется в специальной таблице keyword (см. табл.18).

Таблица 3: Структура записи таблицы method. Справочная информация о локальных архивах

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| method.instrument | метод наблюдений | PRIMARY KEY |
| method.longnameru | русское название из расписания | — |
| method.urldocru | URL русского описания прибора | — |
| method.longnamen | английское название из расписания | — |
| method.urldocen | URL английского описания прибора | — |
| method.telescope | телескоп | — |
| method.focus | фокус телескопа | — |
| method.focuslong | полное название фокуса | — |
| method.obsmode | название режима наблюдений | REFERENCES obsmode |
| — | | ON UPDATE CASCADE |

Таблица 4: Структура записи таблицы telescope. Название телескопа

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| telescope.telescope | телескоп | — |
| telescope.name | вариант названия телескопа | NOT NULL UNIQUE |

Отметим, что в заголовках файлов для локального архива ZMCCD ключевое слово “TELESCOP” может иметь значение “6-м”. Это ошибка, связанная с неправильным заполнением заголовка файла.

С названиями инструментов ситуация такая же, как и с телескопами, поэтому варианты именования приборов, которые встречаются в заголовках наблюдательных файлов, собраны в одну таблицу instrument (см. табл. 5), позволяющую установить их соответствие принятому в поисковой системе названию. В таблице 64 записи.

Отметим, что для локальных архивов LYNX, NES, PFES для правильного определения прибора надо проверять ключевое слово “COMMENT”, в котором уточняется прибор, использовавшийся в наблюдениях.

Таблица detector (см. табл. 6) связывает названия светоприемных устройств, которые обнаруживаются в заголовках наблюдательных файлов, и идентификационный номер, принятый для детектора в поисковой системе. Таблица заполнена после программного анализа заголовков наблюдательных файлов и просмотра полученных уникальных значений. Для примера в таблице 7 представлены ключевые слова и отобранные варианты их значений, выданные программой, анализирующей FITS-заголовки. В таблице 22 записи для 12 светоприемников. Отметим, что название светоприемника может указываться в ключевом слове “INSTRUME” как часть названия наблюдательного прибора.

По таблице obsmode (см. табл. 8) определяются режимы наблюдений на приборах, используемых на телескопах. Представление и структура данных в архивном файле связана с ним. В таблице 9 приведены режимы, которые используются в поисковой системе архива. Если у

Таблица 5: Структура записи таблицы instrument. Астрономический прибор

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| instrument.instrument | прибор | — |
| instrument.longname | название прибора из заголовка файла | NOT NULL UNIQUE |

Таблица 6: Структура записи таблицы detector. Светоприемник

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| detector.detector_id | номер светоприемника | CHECK (detector_id >0 |
| detector.detector | название светоприемника | NOT NULL UNIQUE |

Таблица 7: Ключевые слова FITS-заголовка и примеры значений для определения типа светоприемника

| Лок. архив | Ключевое слово | Формат значения |
|---|---|---|
| LYNX, PFES | — | — |
| NES | DETNAME | “LORAL 2k x 2k” |
| CCD | INSTRUME | “CCD1000 in the PF”, “PMCCD in the PF” |
| CEGS | INSTRUME | “CCD+ECHELLE |
| — | ORIGIN | “PC and CCD K585” |
| IFP, MOFS, MPFS | DETECTOR | “CCD1050x1170”, “TK1024” |
| MSS, SP124 | NSTRUME | “MSSP with CCD1000”, “PMCCD in the Nasmyth-1” |
| RATAN | — | — |
| Scorpio | DETECTOR | “EEV CCD42-40” |
| ZMCCD | INSTRUME | “CCD500 in the CF” |
| — | DETECTOR | “EEV CCD42-40” |
| Z600 | Нет заголовков | — |
| UAGS | INSTRUME | “PMCCD with UAGS”, “TV-channel” |
| — | Instrument | “UAGS+GAD-1” |
| ZMUAGS | Instrument | “ZUAGS+GAD-1” |

Таблица 8: Таблица obsmode. Режим наблюдений

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| obsmode.obsmode | название режима наблюдений | PRIMARY KEY |
| obsmode.descr | описание режима наблюдений | — |

Таблица 9: Режимы наблюдений, принятые в поисковой системе

| Режим наблюдений | Идентификатор |
|---|---|
| echelle spectroscopy | ES |
| direct image | IM |
| interferometry with Fabri-Perot | IFP |
| multi-object spectroscopy | MO |
| multi-slit spectroscopy | MS |
| long slit spectroscopy | LS |
| spectroscopy | SP |
| multi-mode | MM |
| undefined | undf |

прибора несколько режимов наблюдений, как у SCORPIO, то этот параметр имеет значение "MM", и режим наблюдений определяется по еще одной таблице kw_mode (см. табл. 10).

Таблица datatype (см. табл. 11) используется для разделения наблюдательных файлов, помеченных в таблице data (см. табл.19) как "obs", на данные, содержащие наблюдения объектов, и калибровки. В ней собраны сочетания букв в имени файла или в значении ключевого слова, по которым можно определить тип данных. Если тип данных наблюдательного файла можно определить как по имени файла, так и по значению ключевого слова, то в программе анализа FITS-заголовка отдается предпочтение имени файла.

В поисковой системе наблюдательные файлы разделяются на:

- "BS" — байес;
- "DK" — темновой кадр;
- "FF" — плоское поле;
- "ST" — стандарт для калибровки спектральных наблюдений;
- "OBJ" — объект;
- "undf" — тип не определяется.

Таблица 10: Таблица kw_mode. Определение режима многомодовых приборов

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| kw_mode.obsmode | режим наблюдений | REFERENCES obsmode |
| — | — | ON UPDATE CASCADE |
| kw_mode.instrument | инструмент | REFERENCES instrument |
| — | — | ON UPDATE CASCADE |
| kw_mode.pattern | значение ключевого слова | NOT NULL UNIQUE |

Таблица 11: Таблица datatype. Определение типа данных

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| datatype.datatype | тип данных | — |
| datatype.pattern | шаблон для определения типа данных | — |
| datatype.origin | признак, откуда выбрано значение | — |
| datatype.descr | описание | — |

Таблица 12: Таблица person. Персона

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| person.person | имя персоны | PRIMARY KEY |
| person.person_id | идентификатор персоны | NOT NULL UNIQUE |
| person.author | Y - автор, N - нет | — |
| person.observer | Y - наблюдатель, N - нет | — |

### 5.2. Справочные таблицы поисковой системы архива, пополняемые программно

Таблицы person (табл. 12) и person_nick (табл. 13) заполняются значениями, взятыми из текстового файла person.lst. Файл подготавливается программой, анализирующей в FITS-заголовках наблюдательных файлов общего архива значения ключевые слов “OBSERVER” и “AUTHOR”. В этом списке для каждого обнаруженного уникального шаблона фамилии подставляется полная фамилия заявителя программы или наблюдателя. Каждой фамилии присвоен уникальный номер и признаки, отмечающие является ли персона автором и/или наблюдателем, и каждый шаблон получает идентификационный номер, соответствующий фамилии.

После просмотра программным фильтром всех заголовков наблюдательных файлов был сформирован текстовый файл, содержащий 9795 уникальных строковых шаблонов, которые отнесены к предполагаемым именам объектов.

Именем объекта считалась строка, имеющая больше двух символов. Если в такой строке один или два символа, то ей присваивался последовательный отрицательный номер, начиная с -1. При составлении этого списка учитывалось, что в ключевое слово “OBJECT”, по значению которого определяется имя объекта, часто записывается тип экспозиции, например, “FLAT” – плоское поле и т.п. Пустым строкам, и строкам, которые совпадают со значениями таблицы datatype (см. табл. 11), присвоено нулевое значение. Остальные значения получили положительный последовательный номер. Информация об именах объектов сохранена в таблице object (см. табл.14).

По списку предполагаемых объектов (положительный номер) web-сервисом sesame (Schaaff, 2004) из базы данных Simbad/Vizier были извлечены координаты. Из имен, по которым удалось определить объекты и извлечь координаты из Simbad/Vizier, была подготовлена таблица sobject (см. табл. 15), где каждый объект получил идентификационный номер.

Таблица program (см. табл.16) заполняется значениями, взятыми из текстового файла pro-

Таблица 13: Таблица person_nick. Список вариантов записи фамилий

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| person_nick.person_nick | имя персоны в fits-шапке | PRIMARY KEY |
| person_nick.person_id | идентификатор персоны | REFERENCES person; ON UPDATE CASCADE |

Таблица 14: Таблица object. Имена астрономических объектов из FITS-файлов

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| object.object | название объекта из FITS-заголовка | PRIMARY KEY |
| object.object_id | номер объекта | NOT NULL |

Таблица 15: Таблица sobject. Астрономические объекты из Simbad/Vizier

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| sobject.sobject_id | номер объекта | PRIMARY KEY |
| sobject.sobject | название объекта из Simbad/Vizier | NOT NULL |
| sobject.raJ | прямое восхождение J2000 (градусы) | — |
| sobject.decJ | склонение J2000 (градусы) | — |

gram.lst. Файл подготавливается программой, анализирующей ключевое слово "PROGRAM" в FITS-заголовках всех наблюдательных файлов. Данные из таблицы в настоящее время не используются при поиске из web-интерфейса наблюдений по названию программы. Для реализации такого запроса создана таблица schedprog (см. табл.17) с названиями заявленных программ из архива расписаний. Запись таблицы, кроме названия программы на русском и английском языке, включает даты начала и конца сета, а также ключ, идентифицирующий программу.

### 5.3. Связь UCD, ключевых слов FITS-формата с параметрами поисковой системы

Значение атрибута записи таблицы поисковой системы для разных локальных архивов может извлекаться из разных ключевых слов заголовков FITS-файлов. Для программного анализа и извлечения параметров наблюдательных файлов используется таблица keyword (см. табл.18). В таблице каждому ключевому слову, которое встречается в наблюдательных файлах, приведено соответствующее ключевое слово FITS-стандарта, а также дескриптор UCD.

В этой таблице заложена возможность добавления новых параметров для расширения типов запросов к архивным данным. В таблице 729 записей.

### 5.4. Таблицы поисковой системы архива с информацией об архивированных файлах

При просмотре и анализе CD/DVD дисков архива установлено, что из-за ошибок и неполноты параметров, описывающих архивные данные, наиболее надежным для идентификации файлов является дата наблюдений, определяемая из имени каталога с наблюдениями одной ночи. Это учитывалось при разработке поисковой системы, поэтому с занесения данных по дате наблюдений начинается архивирование нового CD/DVD-диска в поисковой системе.

Перед помещением в архив новый диск проверяется и копируется во временный буфер, при этом определяется набор операций над его содержимым (переименование, удаление, компрессирование и т.п.) с тем, чтобы в хранилище архива копия диска соответствовала простым

Таблица 16: Таблица program. Названия наблюдательных программ из FITS-файлов

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| program.program | название программы наблюдений | PRIMARY KEY |
| program.program_id | идентификатор программы наблюдений | NOT NULL |

Таблица 17: Таблица schedprog. Названия наблюдательных программ из расписания

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| schedprog.prog_key | ключ программы наблюдений | PRIMARY KEY |
| schedprog.date1 | дата начала сета наблюдений | — |
| schedprog.date1 | дата окончания сета | — |
| schedprog.onhand | наличие данных в архиве: | — |
| | "Y" – имеются, "N" – отсутствуют | — |
| schedprog.prog_name_r | русское название | — |
| schedprog.prog_name_e | английское название | — |

Таблица 18: Таблица keyword. Связь FITS-ключевых слов локальных архивов, UCD и параметров ИПС

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| keyword.instrument | название инструмента | REFERENCES method ON UPDATE CASCADE |
| keyword.keyword | ключевое слово в FITS-заголовке файлов локального архива | — |
| keyword.fits | соответствующее ключевое слово в FITS-стандарте | — |
| keyword.ucd | UCD дескриптор | — |
| keyword.param | название параметра поисковой системы | — |
| keyword.descr | описание ключевого слова | — |

Таблица 19: Таблица data. Архивированные файлы

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| data.filekey | идентификатор файла | PRIMARY KEY |
| data.datefile | дата получения файла | NOT NULL |
| data.diskname | номер архивного диска | NOT NULL |
| data.catname | каталог, где находится файл | NOT NULL |
| data.filename | имя файла | NOT NULL |
| data.filesize | размер файла (в байтах) | — |
| data.instrument | локальный архив | REFERENCES method; ON UPDATE CASCADE |
| data.filetype | тип файла (obs, log, aux) | — |
| — | (diskname,catname,filename) | UNIQUE |

Таблица 20: Таблица obsparam. Параметры файла (из ключевых слов заголовка файла)

| Название поля | Содержание | Ограничения целостности |
|---|---|---|
| obsparam.filekey | номер файла | REFERENCES data ON UPDATE CASCADE |
| obsparam.instrument | прибор | REFERENCES method ON UPDATE CASCADE |
| obsparam.detector_id | светоприемник | |
| obsparam.datatype | тип данных | |
| obsparam.obsmode | режим наблюдений | |
| obsparam.naxis1 | число пикселей по 1-ой оси | |
| obsparam.naxis2 | число пикселей по 2-ой оси | |
| obsparam.step1 | размер пикселя по 1-ой оси | |
| obsparam.step2 | размер пикселя по 2-ой оси | |
| obsparam.ra | видимое прямое восхождение | |
| obsparam.dec | видимое склонение | |
| obsparam.raJ | прямое восхождение J2000 | |
| obsparam.decJ | склонение J2000 | |
| obsparam.azimuth | азимут | |
| obsparam.zendist | зенитное расстояние | |
| obsparam.object_id | номер объекта по параметру из FITS-заголовка | REFERENCES object ON UPDATE CASCADE |
| obsparam.sobject_id | номер из списка объектов, разрешенных по имени в Simbad/Vizier | REFERENCES sobject ON UPDATE CASCADE |
| obsparam.program_id | номер программы | REFERENCES program ON UPDATE CASCADE |
| obsparam.author_id | список идентификаторов авторов | |
| obsparam.observer_id | список идентификаторов наблюдателей | |

правилам, а именно:

- каталоги на диске содержат данные одной наблюдательной ночи,
- их имена соответствуют дате наблюдений, представленной в следующем формате “YYYYMMDD”,
- вложенные каталоги допускаются, но данные в них относятся к дате каталога верхнего уровня,
- смысловая единица в поисковой системе - файл с наблюдением,
- компрессия данных выполнена алгоритмом bzip2.

После преобразований создается список файлов диска, который используется для наполнения таблицы data. Список содержимого CD/DVD диска и список преобразованного содержимого хранятся в текстовых файлах в специальном каталоге. Эти списки являются частью архива и используются для восстановления информации об отдельном диске или всех дисках в таблице data, обеспечивая один из уровней backup&recovery схемы базы данных.

Структура записи таблицы data представлена в таблице 19. Она является основной для реализации запроса по дате наблюдения, копированию из динамической веб-формы данных, просмотра FITS-заголовка, предпросмотру содержимого файла. В ней содержится ∼240000 записей.

В таблице obsparam (см. табл.19) находятся параметры наблюдательных файлов, извлека-

емые из FITS-заголовков. Это основная таблица для поиска данных по имени объекта, координатам, наблюдательной программе, авторам и наблюдателям. В ней собрана информация о файлах с наблюдениями, которые имеют расширение: fts, fits, mt, а также о части файлов из локальных архивов MPFS и ZMUAGS с расширением tar (это архив из трех файлов, относящихся к одному наблюдению; один из файлов - текстовый, с параметрами наблюдения). Подробнее о заполнении полей записи таблицы obsparam:

• атрибут записи obsparam.filekey является идентификатором файла, который совпадает с таким же полем таблицы data (см. табл.19);

• поле obsparam.instrument содержит название астрономического прибора. Оно определяется сравнением значения параметра, извлеченного из FITS-заголовка файла, со значениями таблицы method (см. табл.3);

• obsparam.detector_id – номер светоприемника. Значение определяется по таблице detector (см. табл.6);

• obsparam.datatype – тип данных в файле (объект, плоское поле и т.д.). Для определения значения этого поля используется справочная таблица datatype (см. табл.11), где собраны сочетания букв в имени файла или в значении ключевого слова, по которым можно определить тип изображения;

• obsparam.obsmode – режим наблюдений (прямые снимки, спектры и т.д.). Значение определяется из справочной таблицы obsmode (см. табл.8, 9), а для приборов с несколькими режимами наблюдений используется еще kw_mode (см. табл.10);

• obsparam.naxis1, obsparam.naxis2 – размер матрицы данных по ключевым словам FITS-заголовка “NAXIS1”, “NAXIS2”. Если “NAXIS2” не имеет значения, то данные являются одномерным массивом;

• obsparam.step1, obsparam.step2 – размер пикселя изображения (в угловых секундах) по ключевым словам FITS-заголовка “STEP1”, “STEP2”. Если эти атрибуты равны единице, то масштаб изображения не определен;

• obsparam.ra, obsparam.dec – небесные координаты объекта на момент наблюдения (центр матрицы).

Эти параметры в заголовок файла передаются из системы управления телескопом (для БТА). Для сервисных файлов также записываются какие-то фиктивные координаты, которые не имеют отношения к объектам. Поиск данных по координатам объекта имеет смысл только для файлов с наблюдениями объектов, поэтому файлы с байесами, темновыми кадрами и т.п. не рассматриваются, но поля записей заполняются значениями координат, которые обнаруживаются в заголовке файлов.

Заметим, что в заголовках некоторых файлов есть дублирующие ключевые слова. К примеру, ключевые слова “RA”, “DEC” и “RA-OBS”, “DEC-OBS” содержат координаты видимого места объекта, только первая пара представляет эти величины в градусах, а вторая – строки с представлением координат в часах, минутах, секундах (см. табл.21). При заполнении записей используется первая попавшаяся пара.

Для локального архива ZMCCD в заголовки файлов пишется азимут и зенитное расстояние.

• obsparam.raJ, obsparam.decJ – координаты, преобразованные в систему каталога FK5 на эпоху 2000.0. Преобразование на стандартную эпоху выполняется программой, разработанной В.С. Шергиным;

• obsparam.object_id – номер объекта из таблицы object (см. табл.14). Имя объекта берется из FITS-заголовка;

• obsparam.sobject_id – номер объекта из таблицы sobject (см. табл.15), где собраны объекты, имена которых разрешаются в Simbad/Vizier. Если имя не разрешается в этих базах данных, то в поле записано пустое значение (“NULL”);

• obsparam.program_id – номер программы из таблицы program (см. табл.16);

Таблица 21: Ключевые слова и форматы значений в FITS-файлах для определения координат объектов

| Лок. архив | Ключевое слово | Формат |
|---|---|---|
| LYNX, PFES | "RA", "DEC" | ddd.dd (градусы) |
| — | "RA-OBS", "DEC-OBS" | "hh mm ss.ss", "±dd mm ss.s" |
| NES | "RA", "DEC" | "hh mm ss.s", "±dd mm ss" |
| — | — | или угл. секунды |
| CCD, MSS | "RA", "DEC" | ddd.dd (градусы) |
| — | "RA-OBS", "DEC-OBS" | "hh mm ss.ss", "±dd mm ss.s" |
| CEGS | "RA", "DEC" | "hhmmss.s", "±ddmmss" |
| IFP, MOFS, MPFS | "RA", "DEC" | "hh mm ss.s", "±dd mm ss" |
| RATAN | "POSTN-RA", "POSTN-DEC" | видимое место (градусы) |
| — | "RA-EPOCH", "DEC-EPOCH" | координаты на эпоху (градусы) |
| Scorpio | "RA", "DEC" | "hh mm ss.s", "±dd mm ss" |
| SP124 | "RA", "DEC" | ddd.dd (градусы) |
| ZMCCD | "AZIMUTH", "ZENDIST" | ddd.dd (градусы) |
| Z600 | — | — |
| UAGS, ZMUAGS | "Right Ascesion", "Declination" | "hh mm ss.ss", "±dd mm ss.s" |
| — | "RA", "DEC" | "hh:mm:ss.ss", "±dd:mm:ss.s" |

Таблица 22: Представление v_disk. Описание архивного CD/DVD-диска

| Название поля | Содержание |
|---|---|
| v_disk.diskname | название диска |
| v_disk.instrument | локальный архив, к которому отнесен диск |
| v_disk.startdate | первая дата наблюдений на диске |
| v_disk.findate | последняя дата наблюдений на диске |
| v_disk.ncatalog | количество каталогов на диске (число ночей) |
| v_disk.nfiles | количество файлов на диске |
| v_disk.volsize | размер диска (в байтах) |

• obsparam.author_id – список номеров авторов, которые были перечислены в ключевом слове "AUTHOR" в FITS-заголовке. Номера, идентифицирующие авторов, определяются по таблице person_nick (см. табл.13);

• obsparam.observer_id – список номеров наблюдателей, которые были перечислены в ключевом слове "OBSERVER" в FITS-заголовке. Номера, идентифицирующие наблюдателей, определяются по таблице person_nick (см. табл.13).

### 5.5. Представления поисковой системы архива

Представление v_disk связывает CD/DVD-диск с локальным архивом (см. табл.22). Записи представления формируются из таблицы data. Заметим, что для диска, на котором записаны данные нескольких локальных архивов, создается столько записей, сколько на нем архивов. Представление используется при динамической генерации web-интерфейса.

Представление v_method (см. табл.23) отображает информацию о локальных архивах в поисковой системе, используется при динамической генерации web-интерфейса.

Представление v_keyword (см. табл.24) связывает ключевые слова FITS-заголовка с параметрами информационно-поисковой системы архива.

Таблица 23: Представление v_method. Методы наблюдений (инструменты)

| Название поля | Содержание |
|---|---|
| v_method.instrument | сокращенное название метода наблюдений (инструмента) |
| v_method.startdate | первая дата наблюдений в методе |
| v_method.findate | последняя дата наблюдений в методе |
| v_method.nfiles | количество файлов в методе |
| v_method.ncatalog | количество каталогов в методе (число ночей) |
| v_method.volsize | общий объем данных в методе (в байтах) |

Таблица 24: Представление v_keyword. Соответствие параметра ИПС ключевому слову FITS-заголовка

| Название поля | Содержание |
|---|---|
| v_keyword.method | сокращенное название метода наблюдений (инструмента) |
| v_keyword.keyword | первая ключевое слово в FITS-заголовке |
| v_ keyword.param | название параметра ИПС |

## 6. Включение архивных данных в инфраструктуру виртуальной обсерватории

Чтобы наблюдения можно было использовать в инфраструктуре виртуальной обсерватории, необходимо чтобы доступ к данным обеспечивался web-сервисами, совместимыми со стандартами IVOA, и данные должны быть готовыми для научных исследований.

### 6.1. О реализации web-сервиса по протоколу SIA к архивным данным

Для извлечения наблюдений из общего архива обсерватории требуются web-сервисы, позволяющие получать из локальных архивов изображения указанного участка неба или спектральные данные для объектов, туда попадающих. Программа, реализующая такой сервис, должна использовать протоколы SIAP (Tody and Plate, 2004) или SSAP (Dolensky and Tody, 2004). Остановимся более подробно на SIA протоколе. Протокол – это набор правил в программной реализации сервиса для клиента, запрашивающего данные, и сервера, организующего их передачу. При получении данных выполняется запрос изображения, подготовка изображения для передачи и собственно передача.

Web-сервис по требованию клиента передает изображение области неба заданного размера. В идеальном случае это некий участок виртуального неба, который реально может состоять из нескольких цифровых изображений, покрывающих эту область, и пользователь не должен беспокоится о стыковке границ отдельных кадров и их калибровке.

SIAP разделяется на четыре режима работы в зависимости от типа изображения и операций с ним. Для архива САО подходит режим, называемый “Pointed Image Archive”, который определяет доступ к коллекциям изображений небольших областей неба, и используется для архивов наблюдений. Остановимся более подробно на его спецификации.

Первое действие – это запрос изображения. Сервис возвращает URL-ссылки на изображения, наиболее подходящие условиям запроса. Ввод данных выполняется как GET-запрос HTTP-протокола. Выглядит это следующим образом:

http:// <server-address> / <path-to-service-program>?[extra-GET-arguments>&[...]]

Сервис передает параметры, определяющие координаты центра и размеры запрашиваемой области неба. Координаты центра задаются в градусной мере в системе координат ICRS (International Celestial Reference System), что соответствует каталогу FK5 на эпоху J2000.0. Размеры

области задаются также в градусах. Сервис должен поддерживать параметр "FORMAT", который отмечает формат или форматы полученных изображений. Они могут быть следующие: fits, html, jpeg, png.

В запросе может передаваться дополнительный параметр "INTERSECT", который определяет, каким образом выбранные изображения должны совпадать с запрашиваемой областью неба. Для архива САО этот параметр опускается, поскольку по умолчанию полагается, что изображение удовлетворяет условию запроса, если частично перекрывается с запрашиваемой областью (INTERSECT=OVERLAY). Размер, масштаб, тип проекции могут использоваться в качестве дополнительных параметров для уточнения запроса.

Результат выдается в виде таблицы в формате VOTable (Ochsenbein et al., 2004) с изображениями, удовлетворяющими условиям запроса. В спецификации определяется, какие элементы VOTable формата являются обязательными. Для описания каждого изображения в таблице используется одна строка с набором параметров в виде UCD-дескрипторов (Derriere et al., 2004). Передается информация, идентифицирующая изображение, координаты и размеры, спектральный диапазон, произведенные действия с изображением. Координаты изображения представляются в стандарте FITS WCS (World Coordinate System) (Greisen and Calabretta, 2002).

Web-сервис на основе SIA-протокола применительно к общему архиву САО требует наличия в поисковой системе следующих параметров для файла наблюдений: даты наблюдения, координат объекта, размера фрейма и угловых размеров пикселя, а, следовательно, еще информации о телескопе, приборе и детекторе, на которых получены данные.

Поскольку параметры изображений, удовлетворяющих запросу, описываются при выводе в таблице UCD-дескрипторами, то для локальных архивов необходимо устанавливать соответствие ключевых слов FITS-заголовка и элементов UCD.

При разработке схемы поисковой системы общего архива обсерватории учитывались требования спецификации SIA-протокола, а в результате наполнения таблиц имеется весь набор необходимых параметров для реализации web-сервиса, совместимого со стандартами IVOA, для доступа к архивным данным.

Из рассмотреных нами программных средств для организации web-сервисов по протоколам IVOA к коллекциям FITS-файлов только программный пакет генерации баз данных Saada (Nguyen et al., 2006) поддерживает такие сервисы на стороне сервера. Пакет позволяет создавать из однородной коллекции FITS-файлов с изображениями или спектрами базу данных на основе СУБД PostgreSQL и обеспечивает доступ к наблюдениям и выдачу результатов по протоколам IVOA, а также запросы на языке запросов SaadaQL. Однако, применение его к архивным данным требует дополнительной работы по коррекции и дополнению параметров наблюдательных файлов или пропущенных значений, так, например, поиск наблюдений по заданным координатам можно применить к части архивных файлов с наблюдениями объектов (не в каждом имеются координаты).

### 6.2. Подготовка "science-ready" данных

Общий архив САО - это коллекция "сырых" данных, и даже их первичная обработка, связанная с редукцией инструментальных ошибок, является сложной задачей. Не только погодные условия, но и характеристики светоприемников меняются в течение ночи, поэтому только при наличии в локальных архивах для каждого сета хороших по качеству сервисных файлов (плоских полей и т.п.) можно говорить о динамической обработке данных по запросу. Судя по имеющимся астрономическим программным пакетам, таким как, например, PLEINPOT (Chilingarian et al., 2005), можно адаптировать эти средства для обработки по запросу архивных данных. Поисковая система архива предоставляет затребованное наблюдение с соответствующими сервисными файлами для обработки, а программный пакет позволяет выполнить стан-

дартную последовательность действий. Сложнее всего в нашем случае обеспечить обработку необходимыми сервисными файлами и параметрами наблюдений, которые часто отсутствуют в локальных архивах.

Возможны разные варианты подготовки архива "сырых" наблюдений и преобразования его в данные готовые для научного анализа ("science-ready"):

1. из-за того, что данные требуются ограниченному кругу лиц, оставить их необработанными и направить усилия на обеспечение полноты и корректности параметров, описывающих наблюдения. Часть пропущенных параметров можно восстановить программно, так, например, если в заголовке файла указано имя объекта, то имеется возможность дополнить координаты. Сравнение записей в журналах наблюдений с записями в таблицах позволит также дополнить или скорректировать частично информацию о наблюдениях. Если автоматическая коррекция параметров невозможна, то требуется участие авторов наблюдений в исправлении ошибок посредством соответствующего интерфейса, а также авторизации и подтверждения прав пользователя на такие операции с данными;

2. добавить в общий архив имеющиеся у пользователей обработанные данные, обеспечив соответствующие изменения в web-интерфейсе архива. Добавление новых коллекций в архив не требует больших изменений в поисковой системе;

3. предоставить программное средство пользователю для индивидуальной обработки наблюдений. Одним из примеров такой обработки служит пакет pleinpot, на основе которого производится редукция данных из архива наблюдений HyperLEDA (http://leda.univ-lyon1.fr/11/development.html);

4. реализовать автоматическую обработку данных по запросу. Для всего общего архива наблюдений обсерватории маловероятно организовать автоматическую редукцию данных.

## 7. Заключение

Архивная система должна обеспечивать хранение и сохранность информации, доступ к данным и добавление новых коллекций данных. Этими положениями мы руководствовались при разработке архива.

Для архива обеспечивается нескольких уровней хранения данных: две копии каждого CD/DVD диска и образ дисков в специальной области хранения на файловом сервере обсерватории, хранилище архивированных дисков, а также информация в текстовых файлах с перечнем файлов на каждом архивном диске.

Сохранность данных поддерживается постоянным сопровождением архивной системы, что включает: администрирование операционной системы сервера и системы управления базами данных, верификацию и добавление новых дисков, восстановление архивной системы или ее части (хранилище файлов, поисковая система) при аварийных ситуациях, перенос части архивных данных в новое место расположения (раз в 3-4 года), перезапись (с верификацией) CD/DVD дисков на новый тип носителя (раз в 5-7 лет).

В архивной системе нет ограничений на формат файлов, поэтому при добавлении новых локальных архивов необходимо, чтобы данные наблюдательной ночи хранились в одном каталоге и в название каталога с наблюдениями одной ночи входила дата. Если эти условия выполняются, то добавление нового локального архива не составляет особых трудностей. В ряд таблиц архивной системы добавляются записи, фиксирующие информацию о новом локальном архиве. Вносятся изменения в программы, обеспечивающие Интернет-доступ к данным.

Организован открытый доступ к наблюдениям в соответствии с Положением об архиве наблюдений обсерватории, отвечающий положению резолюции МАС об открытом Интернет-доступе к архивированным наблюдениям, полученным в обсерваториях, финансируемых из государственного бюджета.

**Благодарности.** Работа выполнена при поддержке РФФИ (грант №07-07-00415).

## Список литературы

Билдинг (Building the Framework ...) Building the Framework for the National Virtual Observatory. NSF Cooperative Agreement AST0122449, Quarterly Report, April-June 2004, 44 (2004)
Бартунов О., http://www.sai.msu.su/ megera/postgres/talks/what-is-postgresql.html (2005)
Валтон и др. (Walton, N. A. et al.), Astronomy & Geophysics, 47, 3.22 (2006)
Велс и др. (Wells, D. C., et al.), A&AS, 44, 363 (1981)
Верходанов и др., Препринт САО РАН. СПбФ, 89СПб,18-30 (1993)
Вильямс и др. (Williams R., et al.), http://www.ivoa.net/twiki/bin/view/IVOA/IvoaArchitecture (2004)
Вильямс и др. (Williams R., et al.), http://www.ivoa.net/internal/IVOA/ConeSearch/ConeSearch-20070628.html (2007)
Витковский и др., Сообщения САО, 59, 60 (1988)
Витковский и др. (Vitkovskij V., et al.), Baltic Astronomy, 9, 578 (2000)
Грабер М., Лори-пресс (2003)
Грабер М., ЛОРИ, 375с (1996)
Грейсен и Калабрета (Greisen E.W. and Calabretta M.R.), A&A, 395, 1061 (2002)
Дерье и др. (Derriere S., et al.), http://www.ivoa.net/Documents/UCD/WD-UCD-20040426.html (2004)
Доленски и Тоди (Dolensky M., Tody D.), SPIE, 5493-47 (2004)
ИВОА СкайНоуд (IVOA SkyNode ...) IVOA SkyNode Interface, http://www.ivoa.net/internal/IVOA/IvoaVOQL/SkyNodeInterface-0.7.4.pdf (2004)
Исуда и др. (Ysuda N., et al.), ASP Conf. Ser., 30 (2004)
Камбреси и др. (Cambresy, L., et al.), http://www.ivoa.net/twiki/bin/view/IVOA/IvoaSemantics/WD_2007-02-19.pdf (2007)
Клемен и др. (Clement, L., et al.) OASIS UDDI Specification TC, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=uddi-spec (2003)
Лемсон и др. (Lemson, G., et al.), http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel/ DomainModel0.9.1.doc (2003)
Льюс и др. (Louys, M., et al.), http://www.ivoa.net/Documents/latest/CharacterisationDM.html (2007)
МакДауел и др. (McDowell, J., et al.), http://www.ivoa.net/Documents/latest/DMObs.html (2005)
Нгуен и др. (Nguyen et al.), ASP Conf. Ser., 351, 15 (2006)
МакДауел и др. (McDowell, J., et al.), http://www.ivoa.net/twiki/bin/view/IVOA/IVOADMQuantityWP/qty23.pdf (2004)
Оксенбайн и др. (Ochsenbein, F., et al.), ASP Conf. Ser., 216, 83 (2000)
Оксенбайн и др. (Ochsenbein, F., et al.), http://cdsweb.-strasbg.fr/doc/VOTable/v1.09 (2004)
Паблик (Public Access ...) Public Access to Astronomical Archives. The Resolution of 5 Commission of IAU http://www.atnf.csiro.au/people/rnorris/WGAD/Resolution.htm (2003)
Плейн и др. (Plane R., et al.), ASP Conf. Series, 30 (2004)
Ротс (Rots, A.), http://www.ivoa.net/Documents/latest/STC-Model.html (2007)
Салгадо и др. (Salgado, J., et al.), www.ivoa.net/forum/dm/att-1086/SLAP_v0.2_29_Dec_2005.pdf (2005)
Тейлор (Taylor, M.), http://www.star.bris.ac.uk/ mbt/topcat/sun253/index.html (2007)
Тоди и Плейт (Tody D., Plate R.), http://www.ivoa.net/Documents/WD/SIA/sia-20040524.html (2004)
Тоди (Tody, D.), http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDAL/TAP-Analysis.pdf (2007)
Ферник и др. (Fernique, P., et al.), ASP Conf. Ser., 145 (1998)
Хессман и др. (Hessman, F., et al.), http://www.ivoa.net/internal/IVOA/ IvoaSemantics/VOcabularies-20070903.htm (2007)
Чилингарян и др. (Chilingarian, I., et al.), ASP Conf. Ser., 347, 385 (2005)
Шааф (Schaaf, A.), ASP Conf. Ser., 314, 327 (2004)
Штебе и др. (Stebe, A., et al.), http://www.ivoa.net/twiki/bin/view/IVOA/ IvoaDAL/TAP-Draft.pdf (2007)

## 8. Приложение А. Перечень программ, обеспечивающих работу с поисковой системой архива наблюдений

В комплекс программ, обеспечивающих функционирование информационно-поисковой системы архива обсерватории, входят скриптовые программы на языке программирования Perl (∼5900строк) и командной оболочки bash (∼1700строк), а также SQL-программы для создания схемы таблиц и наполнения таблиц с постоянным содержимым (∼3100строк):

- fetch.pl – web-интерфейс пользователя для реализации запросов к наблюдательным данным. Программа формирует динамический web-интерфейс к общему архиву САО, включающий названия локальных архивов, диапазон дат, формы для реализации запросов и результатов поиска;
- fetchadm_res.pl – web-интерфейс администратора. Вывод результатов запросов в расширенном формате с дополнительной служебной информацией для администратора архива;
- input_cd.pl – занесение содержимого CD/DVD-диска в архив (в основную или отладочную базу данных);
- input_cdD.pl – занесение диска в архив (в основную или отладочную базу данных) со сменой названия каталогов из форматов YMDD, YYMMDD в формат YYYYMMDD;
- delete_cd.pl – удаление диска из из архива (из основной или отладочной базы данных);
- input_md – занесение всех дисков указанного локального архива в ИПС;
- delete_md – удаление всех дисков указанного локального архива из ИПС;
- input_all – занесение всех дисков в архив;
- getobs.pl – получить данные сета из расписания наблюдений по дате и прибору;
- input_hd.pl – анализ FITS-заголовка файла и занесение данных в архив;
- view.pl – просмотр содержимого заданной таблицы в основной или отладочной базе данных;
- a_echelle.pl – подготовка оптического диска локального архива ECHELLE для внесения в архив, при этом уточняется по FITS-заголовкам принадлежность файла локальным архивам (LYNX, NES или PFES) и переименовываются каталоги (для NES и PFES);
- a_lynx – подготовка дисков локального архива LYNX для внесения в архив: перевод файлов в формате bdf в FITS, формирование текстового файла со списком файлов на диске, уточнение принадлежности к локальным архивам (LYNX, NES или PFES);
- to_oasis – удаленное копирование CD/DVD-диска с текущей машины на сервер базы данных;
- rmcd – удаление CD/DVD-диска из указанного локального архива;
- search – поиск образца (суффикса в имени файла) в заданном локальном архиве или по всему архиву;
- a_bdf – подготовка (перевод файлов в формате системы MIDAS в FITS-формат) заданного диска (локальные архивы CCD, LYNX, SP124, MSS) для внесения в архив;
- a_step1 – поиск нескомпрессированных файлов на CD/DVD-диске и их компрессия;
- a_chmod – поменять режимы доступа к файлам CD/DVD-диска только на чтение (код доступа для каталогов – 555, для обычных файлов – 444);
- cd2arch – помещение подготовленной копии CD/DVD-диска в архив;
- a_rename – переименование каталогов и файлов для указанного CD/DVD-диска;
- a_delete – уничтожение каталогов и/или файлов для указанного CD/DVD-диска;
- a_tar – раскомпрессировать tar-архив, разархивировать на отдельные файлы и скомпрессировать их для указанного CD/DVD-диска локального архива SCORPIO;
- a_unzip – раскомпрессировать ZIP-файлы (unzip) и заново скомпрессировать командой bzip2;
- gethd.pl – вывод в html-формате содержимого FITS-заголовка;

• gethdtar.pl – разархивировать tar-архив, выбрать файл с расширением dsv (локальные архивы MPFS,UAGS,ZMUAGS) или hdr (локальный архив Z600) и вывести его содержимое в html-формате;

• keywords.pl – создание списка всех ключевых слов и их значений из заголовков файлов в FITS-формате для указанного CD/DVD-диска и локального архива;

• pop_mk.pl – создание списка всех значений ключевых слов: AUTHOR+OBSERVER, OBJECT, PROG-ID. При перечислении авторов и наблюдателей списком выполняется разбиение по разделителям (пробел, запятая, амперсанд) и отбрасываются инициалы;

• instrum_mk.pl – создание списка всех значений ключевых слов INSTRUME и COMMENT для локального архива LYNX;

• a_del_ren - подготовка CD/DVD-диска для внесения в архив: удаление ошибочных файлов, переименование каталогов из lYMDD в YMDD;

• person_tab.pl - заполнение таблиц person, person_nick значениями, взятыми из файла person.lst. Структура записи person.lst: значение ключевого слова OBSERVER или AUTHOR в квадратных скобках, один или несколько пробелов полная фамилия в квадратных скобках. Таблица person: каждой фамилии программно присваивается уникальный номер и признаки автора и/или наблюдателя. Таблица person_nick: каждый nick получает номер соответствующей ему фамилии и переводится в верхний регистр;

• program_tab.pl - заполнение таблицы program значениями, взятыми из файла program.lst. Структура записи program.lst: значение FITS – PROG-ID в квадратных скобках. Названия программ переводятся в верхний регистр. Каждой программе программно присваивается уникальный номер (id). Программы, названия которых $< 3$ символов считаются unknown и им присваивается номер 0;

• to_base1_Data - удаленно копировать каталог с содержимым CD/DVD-диска с текущей машины на файловый сервер с сохранением атрибутов файлов;

• a_makels - создать текстовый файл со списком файлов помещаемого в архив CD/DVD-диска;

• a_bdf_ren - подготовка CD/DVD-диска для внесения в архив: перевод файлов в формате системы MIDAS в FITS-формат и перенос вложенных каталогов уровнем выше по дереву каталогов;

• a_date.pl - переименование каталогов с названиями в формате YMDD, YYMMDD в YYYYMMDD;

• a_help - помощь по внесению CD/DVD-диска в архив и в базу данных: анализ метки диска на корректность, просмотр журнального файла, занесение в отладочную или основную базу данных, добавление записи о диске в input_md и input_all и его копирование на файловый сервер;

• tst_metka.pl - синтаксический разбор метки CD/DVD-диска и проверка на допустимые значения элементов метки

• cdatetab.sql, crecoparam.sql, credatatype.sql, keywordic.sql – создание и наполнение таблиц схемы базы данных.