Качество астрономических данных в контексте интеграции ресурсов в виртуальную обсерваторию

© Желенкова О.П., Витковский В.В, Пляскина Т.А.

Специальная астрофизическая обсерватория РАН
zhe@sao.ru

Аннотация

Для виртуальной обсерватории (ВО) – web-инфраструктуры, объединяющей многочисленные астрономические цифровые коллекции нужны готовые для научного анализа (science ready) данные. Постоянно растущие объемы информации, интеграция ресурсов в ВО требуют целостного подхода в обеспечения качества данных.

1 Введение

В астрономии, как и во многих других науках, происходит революция в данных, с одной стороны предоставляющая ученым совершенно новые возможности и с другой стороны добавляющая проблем, которые до этого не возникали. Из-за привычной и принятой в астрономии традиции свободного распространения данных ее можно рассматривать как некий тренировочный плацдарм по опробованию информационных технологий для повышения эффективности научных исследований. Пока организация и архивирование, как терабайтных астрономических баз данных, так и ресурсов отдельных обсерваторий все еще остается достаточно бессистемной, чтобы позволить ученому при обращении к ним не вникать в форматы или добывать информацию о структуре и качестве обработки востребованных данных. Из-за того, что методы научных исследований в астрономии часто связаны со сравнением и объединением различных данных, то важно исследователю иметь доступ ко всей имеющейся информации об отдельном объекте, классе объектов или явлении. Но ситуация сейчас далека от желаемой. Необходимые данные, публикуемые в журналах, часто не помещаются в центры данных для коллективного использования. Большое количество информации в обсерваториях недоступно для пользования. В астрономии нужна схема руководящих принципов в контроле и организации данных [1].

Качество данных никогда не было в астрономии забытой темой. Постоянно растущие объемы данных подталкивают к целостному осмыслению этой проблемы. Немаловажным фактором является также то, что астрономические данные не теряют со временем своей научной значимости.

В этом сообщении попытаемся рассмотреть вопросы, связанные с темой качества астрономических, в широком смысле слова, с разных сторон, - целостности, достоверности и погрешности измерений физических величин. Каждый из этих вопросов требует своего подхода для оценки состояния данных, возможного исправления промахов и ошибок, а также маркировки целостности, сохранности, погрешности измерений, полноты и информативности описания ресурса. Обеспечение качества неотделимо от организации и менеджмента данных.

Нижний уровень этой проблемы, – аппаратный, связан с физическим хранением данных, с миграцией данных с одного типа носителя на другой. Следующий уровень – логическое представление данных. При миграции между логическими форматами представления данных, даже между версиями одного формата, могут возникнуть неоднозначности и ошибки, которые приводят к некорректному представлению данных в программных средствах, используемых для анализа.

При реализации сервисов к данным, и не только для этого, необходимо полное семантическое описание данных и достоверные значения параметров, как для одной коллекции, так и для

одного файла. При сравнении данных разных спектральных диапазонов требуется совместимость описаний параметров для корректного преобразования физических величин.

И наконец, для ВО нужны готовые для научного анализа данные (science ready), то есть такие данные, в которых исправлены инструментальные ошибки, выполнены привязка и переход от относительных значений к реальным физическим величинам. Нельзя сравнивать и анализировать недостоверные, плохо привязанные к точкам отсчета физические величины. Невозможно получить научный результат, сравнивая две цифровые коллекции и не имея информации о погрешностях измерений физических величин. Данные переводятся в некое многомерное физическое пространство, они освобождаются от искажающих сигнал факторов: влияния погодных условий, аппаратных функций инструмента и приемного устройства, поглощения в атмосфере и межзвездной среде и так далее. Значения физических величин, полученные из таких данных можно рассматривать как параметры однородного пространства, описывающего свойства небесного объекта, и приступать к сравнительному изучению.

2 Текущее состояние астрономических ресурсов

Астрономия находится в первых рядах среди других научных сообществ в создании инфраструктуры, реализующей связи между разнородными распределенными системами. Главной организацией, которая направляет, координирует деятельность мирового астрономического сообщества, является Международный астрономический союз (International Astronomical Union, IAU), объединяющий ученых из 70 стран. Об основных направлениях деятельности IAU можно более подробно узнать на сайте http://www.iau.org. Одно из подразделений IAU, Пятая комиссия “Documentation and Astronomical Data”, занимается вопросами документирования, стандартизации и менеджмента астрономических данных. Она включает в свою структуру следующие рабочие группы (РГ) - Astronomical Data, FITS, Libraries, Designations, Virtual Observatories, Data Centers & Networks, Preservation & Digitization of Photographic Plates. Деятельность этих подразделений связана, так или иначе, с контролем, управлением и обеспечением качества данных, выработкой для этого спецификаций, стандартов, протоколов [2].

В конце 20 - начале 21 века активное внедрение информационных технологий в астрономические исследования привели сначала к появлению концепции, а затем к созданию инфраструктуры виртуальной обсерватории. Основная идея ВО состоит в том, чтобы реализовать прозрачный доступ к распределенным данным, простую и эффективную работу с ними так, как если бы эти данные находились на компьютере пользователя. Все архивы наблюдений, хранящие в себе еще множество несделанных открытий, должны понимать один и тот же язык запросов, иметь единый интерфейс для доступа к данным, полученные данные можно проанализировать одними и теми же программными средствами. Концепцию ВО можно рассматривать как пример проблемно-ориентированного грида данных. Однако, делается еще один шаг вперед, так как предполагается не просто доступ к данным, но их обработка и возвращение результата, что является существенным для получения реального научного результата. На текущий момент такой анализ возможен только после копирования необходимой информации на компьютер пользователя. ВО в ближайшем будущем позволит использовать веб-сервисы для проведения вычислений с хранимой в центрах данных информацией. Вычислительные сервисы должны быть стандартизованы для совместимости операций по многим архивам. При глобализации астрономических исследований требуется решение ряд задач, в которых качество данных является важным аспектом.

Координацию усилий астрономического сообщества в разработке и развитии инфраструктуры ВО выполняет созданный в 2002 году международный альянс IVOA (International Virtual Observatory Alliance), объединяющий 15 национальных проектов (http://www.ivoa.net).

С момента появления проектов ВО и организации IVOA в астрономическом сообществе велось много дискуссий по поводу достоверности и надежности научных результатов при совместном использовании разнообразных и разнородных данных, получаемых из астрономических ресурсов. Прогресс в разработке ВО-совместимых программных инструментов, веб-сервисов, обеспечивающих доступ и анализ данных показал, что опасения по этому поводу необоснованны. Эти работы скорее, наоборот, ускорят решение задачи обеспечения и контроля качества данных.

В IVOA работает несколько РГ, затрагивающих эту проблему с разных сторон, это - Resource Registry, Content Descriptor (UCD), Data Curation & Preservation.

Вернемся к имеющимся и планируемым при введении новых проектов объемам данных в астрономии. Один из самых успешных проектов последних лет в области проведения цифровых обзоров – Слоановский цифровой обзор неба (Sloan Digital Sky Survey, SDSS, http://www.sdss.org). Терабайтный объем данных, новые подходы в технике наблюдений, полностью автоматизированная обработка данных, применение новых информационных технологий в хранении и доступе к данным сделали этот проект прототипом концепции ВО.

В статье [5] двух ведущих специалистов этого проекта Шалои и Грея рассматриваются проблемы хранения, доступа и анализа таких объемов информации. Традиционный подход ученых при анализе результатов наблюдений или экспериментов состоит в хранении данных и процедур обработки на своих рабочих компьютерах с тем, чтобы можно было повторно воспроизвести полученный результат. Такой подход оправдывает себя тогда, когда персональный компьютер справляется с количеством данных и объемом вычислений. Но количество данных удваивается ежегодно в большинстве областей современной науки, и их анализ становится все более сложным. Популярные программные средства для анализа данных не приспособлены для работы с миллионами записей. Из-за объема данных существенно возрастает сложность извлечения знания. Исследователь должен прилагать значительные усилия для организации, сортировки и обработки данных. Анализировать терабайты данных – это проблема, но наборы данных размером в петабайты уже на горизонте. В ответ на такой вал данных систематическое использование баз данных становится неотъемлемой частью научного процесса. Для этого требуются новые стандарты для обмена данными, а также семантический словарь, который предлагает контролируемый лексический перечень астрономических терминов. Создатели и организации, публикующие данные, должны обеспечивать достоверность данных, автоматически предоставлять информацию об источнике.

Совместное сотрудничество сотен ученых с помощью Интернета ставит вопросы о стандартах по разделяемому использованию данных. Слишком много сил расходуется при переформатировании данных из одного формата в другой. Стандарты необходимы на нескольких уровнях [1]:

- при форматировании данных; чтобы одна группа с легкостью могла читать и понимать данные другой группы;
- в семантике; чтобы термин, используемый одной группой, легко переводился в другой, аналогичный по семантическому значению, без искажения смысла;
- в процедурах обработки; чтобы отдельные этапы можно было выполнять в Интернете с возможностью воспроизведения этапов обработки другими исследователями.

Базы данных имеют механизмы и обеспечивают средства для организации работы с большими наборами данных, эффективного поиска, сортировки.

В 2003 году Генеральная Ассамблея IAU приняла резолюцию об открытом доступе к архивированным данным [4]. Данные, полученные обсерваториями, финансируемыми из государственного бюджета, после обоснованного периода права собственности, при котором данные доступны только авторам, должны быть помещены в архив, где Интернет-доступ к ним открывается для всех желающих. На сколько это возможно, данные должны сопровождаться соответствующими метаданными или программными средствами, для того, чтобы можно было извлекать из них научную информацию для анализа. Такие данные не уже являются предметом интеллектуальной собственности.

Чтобы открыть веб-доступ к отдельному астрономическому каталогу, обзору, архиву, надо, прежде всего, создать базу данных, обеспечить сервисы для работы с ней и ее сопровождение. Пока нет какой-либо стратегии, которая связывала бы вместе эти виды деятельности, обеспечивала бы политику для управления астрономическими данными. В результате этого [1]:

- нет единого подхода в обеспечение сохранности и распространении данных;
- пока одни группы внедряют профессиональный подход к менеджменту данных, другие откладывают его на потом, третьи отвергают это полностью, таким образом, в астрономии теряются данные;
- слабая координация между астрономией и другими дисциплинами в том, какие данные

могли бы использоваться в исследованиях.

На предстоящей 26 Генеральной ассамблее IAU в августе 2006 года запланировано рассмотрение вопросов, связанных с разработкой протоколов долговременного курирования, сохранности и миграции данных (между форматами и носителями), включая маркировку качества и правильности данных [2].

3 Сервисы IVOA и обеспечение качества данных

На конференции IVOA в 2003 году были определены шесть основных технических инициатив, необходимых для достижения научных целей ВО [2]. Они определены из потребностей общей архитектуры ВО, которая должна обеспечивать интероперабельный и прозрачный доступ к данным. Перечислим эти разработки:

- *Регистры*. Регистры отвечают за сбор метаданных об астрономических ресурсах и информационных сервисах в базу данных и обеспечивают запросы по их поиску.
- *Модели данных*. В качестве семантического стандарта для астрономии необходима доменная модель данных, которая позволит разрабатывать программное обеспечение, работающее с множеством вариантов представления данных без модификации структур источников данных.
- *Uniform Content Descriptors (UCDs).* Для установления семантической связи между наименованиями и понятиями IVOA поддерживает контролируемый словарь дескрипторов.
- *Data Access Layer (DAL).* DAL обеспечивает механизм стандартного доступа к распределенным астрономическим данным.
- *VO Query Language*. Для работы с многочисленными распределенными наборами данных требуется стандартный язык запросов. Хотя SQL можно использовать для запросов к большинству современных астрономических баз данных, но астрономическая специфика требует расширения возможностей языка запросов.
- *Grid & Web Services*. ВО – это грид сервисов, где программные средства и коллекции данных располагаются вместе на одном узле. Дальнейшие развитие инфраструктуры - реализация методов асинхронных сообщений, авторизация подписью и управление потоками работ. Совместная работа сервисов с данными требует управления памятью на каждом узле ВО (VOSpace). Эта расширяемая, видимая в Интернете память выделяется автоматическим процессам и пользователям для обмена между процессами и совместной работы.
- *VOTable.* Этот формат для астрономических таблиц был первым соглашением, принятым IVOA. В VOTable используется индустриального стандарта XML и наследуется опыт разработок FITS и CDS Astrores форматов. Он используется в сервисах ConeSearch, SIAP, SSAP для представления результатов запросов.

Разработкой интерфейсов, протоколов, спецификаций занимаются рабочие группы альянса. Теперь попытаемся проследить, как предполагается фиксировать, отображать, передавать и контролировать информацию о качестве данных в инфраструктуре ВО. Начнем с рекомендации по описанию ресурса в регистре ВО, разработанной рабочей группой Resource Metadata (RM WG)и принятой IVOA [7]. Регистры являются распределенными источниками информации о ресурсах и сервисах. Для определения механизмов реализации регистров RM WG было рассмотрено большое число индустриальных стандартов, включая Open Archive Initiative (OAI), разработанную для цифровых библиотек. Для описания метаданных регистров используются определения Dublin Core. Все ресурсы описываются метаданными общего плана. Они включают:

- идентификацию метаданных - имя и идентификатор ресурса;
- сопровождение ресурса, в котором описывается, кто поддерживает ресурс и его наличие (версия, дата релиза);
- метаданные контента, в которых указывается тип информации (тип данных, область неба, занимаемая данными, спектральный диапазон и т.д.).

Метаданные ресурса составляют “yellow pages” астрономической информации. Они являются аналогом UDDI (Universal Description, Discovery and Integration) для веб-сервисов или GLU, используемого в CDS. GLU – программный пакет

для разрешения URL-ссылок на астрономические ресурсы [8]. В рекомендацию входит и экспертная оценка качества данных. Коротко о положениях этого пункта рекомендации [7].

Пользователям ресурсов ВО необходимо иметь информацию о качестве данных. Коллекция данных может не иметь этой оценки. Полнота и согласованность самого описания уже является индикатором качества ресурса., предполагается развитие и рекомендация его повсеместного использования в инфраструктуре ВО. Элементы метаданных содержат экспертную оценку общего качества ресурса и погрешностей в привязках к реальным осям многомерного пространства. Вот эти параметры:

- *DataQuality* – общая оценка целостности, согласованности и уровня документированности, касающегося оценок погрешности и калибровочных процедур для данных ресурса. Предлагается 3 уровня градации и кодировка неизвестных и недокументированных случаев:

  A – полная калибровка данных, имеется полная документация, пригодны для профессиональных исследований (science-ready);

  B – имеются калибровки и описания, но качество калибровок не согласовывается. Пользователям необходимо проверять данные и провести повторную калибровку;

  C - некалиброванные данные;

  U – качество данных неизвестно. Если описание ресурса не содержит оценки качества данных, то им присваивается класс U.

- *ResourceValidationLevel* – числовой параметр, описывающий качество описания и интерфейс и отмечающий степень достоверности ресурса при включении его в ВО приложение или при использовании в научных задачах. Этих уровней достоверности предполагается 5, от 0 до 4. Если ресурс не описан или его описание не соответствует стандарту, то он имеет уровень достоверности равный 0. 1 и 2 – описание соответствует стандарту, сервис демонстрирует функциональную совместимость со стандартом, при запросе выдает документ без ошибок.. 3 и 4 – описание ресурса проверяется экспертом, установлены параметры качества, используется ВО-приложением. В отличие от других описаний ресурса значения ResourceValidationLevel чаще всего будут устанавливаться не провайдером данных, а администратором регистра. В информационной инфраструктуре, когда запись ресурса может существовать в нескольких регистрах, каждый экземпляр записи может иметь свое значение, в зависимости от практики и стандартов качества, принятых в регистре. Уровни 0, 1, и 2 определены так, что могут назначаться автоматически программным агентом. Уровни 3 и 4 требуют участия эксперта при установке значений.
- *ResourceValidatedBy* - IVOA идентификатор [9] для регистра или организации, установивших значение для ResourceValidationLevel.
- *Uncertainty.Photometric* - погрешность фотометрических измерений (в Янских).
- *Uncertainty.Spatial* – погрешность астрометрических или позиционных измерений данных (в градусах).
- *Uncertainty.Spectral* - погрешность определения длин волн (в метрах).
- *Uncertainty.Temporal* – погрешность шкалы времени.

Наиболее полно описание ресурсов, соответствующее этой рекомендации, поддерживается существующими регистрами NVO [10]. Регистры NVO позволяют без ограничения любому пользователю выполнять регистрацию своего ресурса. В регистре имеется более 10000 записей. Но в них часто отсутствуют значения для большей части метаданных, описывающих ресурс. В Страсбургском центре данных действует похожий механизм регистрации и разрешения URL-ссылок на ресурсы – GLU [8], который был разработан исторически раньше, чем принята рекомендация [7]. Используется он в CDS, регистрация выполняется только сотрудниками центра. В GLU-регистре указывается версия данных, никаких дополнительных оценок правильности и качества данных не предоставляется. Для информационной структуры AstroGrid действует свой регистр с отличающимся набором описаний. Имеются интерфейсы для взаимодействия трех типов регистров.

Рабочая группа IVOA Data Models (DM

WG) работает по нескольким направлениям, связанным с семантическим представлением данных в астрономии, что включает разработку доменной модели данных для астрономии, UCDs, участие в разработке моделей данных для DAL.

Хотя астрономическое сообщество использует общий формат данных - FITS стандарт, существует много вариантов представления метаданных в FITS файлах, и множество способов представления данных, связанных с небесными объектами (например, спектры, шкала длин волн и их погрешности). FITS не является семантическим стандартом, поскольку его синтаксис не позволяет однозначно интерпретировать и определять параметры файлов. В качестве семантического стандарта разрабатывается доменная модель данных [11]. Она еще находится на начальном этапе работы, поэтому вопросы, связанные с качеством данных более подробно проработаны в компонентных моделях, являющихся ее составными частями. В текущей разработке находятся несколько компонентных моделей, а именно: наблюдения (Observation), представления физических величин (Quantity), пространство-время (Space-Time Metadata). В Observation [12] описываются астрономические наблюдения, где среди прочих параметров определяется астрометрическая, фотометрическая, временная точность наблюдений. Уточнение описания характеристик ошибок выполняется в модели данных Quantity. В ней определяются ошибки (абсолютные относительные, систематические, случайные), цифровой диапазон значений, размеры пикселя, разрешение инструмента [13].

Рабочая группа IVOA Semantics развивает технологии семантического описания астрономических данных для обеспечения интероперабельности ВО систем [14]. Деятельность группы направлена на определение значений и интерпретацию слов, предложений или других языковых форм в астрономическом контексте. Сюда относятся описания астрономических объектов, типов данных, астрономических понятий и явлений. В рабочей группе рассматриваются взаимосвязи между словами, символами и понятиями, значения представлений, использование естественного языка в астрономии, включая запросы, переводы, интернационализацию интерфейсов. Начиная с уже имеющихся списков астрономических слов, объектов и понятий, в частности с Unified Content Descriptor (UCD), необходимо определить основные понятия, их экземпляры и отношения между ними. В Страсбургский центр звездных данных (Centre de Données astronomiques de Strasbourg, CDS), где имеется крупнейшая коллекция астрономических каталогов и таблиц, разработан словарь дескрипторов UCDs для того, чтобы получать семантическое значение из множества наименований колонок. В словарь вошло около 1500 уникальных описателей типов содержимого колонок из имеющихся десятков тысяч наименований. Усовершенствованный стандарт для дескрипторов UCD1+[6], создает более согласованный и расширяемый набор элементов, с меньшим количеством элементарных UCDs. Рабочая группа обеспечивает:

- сопровождение UCDs;
- стандартный словарь IVOA;
- изучение онтологий.

Основная деятельность группы состоит в определении основных элементов стандартного словаря для всех областей деятельности IVOA. Все протоколы и интерфейсы IVOA используют UCDs для установления смысловых отношений между параметрами запросов и метаданных.

Задача рабочей группы DAL (DAL WG) состоит в определении стандартов ВО для дистанционного доступа к данным. Клиентские программы, совместимые с этими стандартами, смогут использовать сервисы для доступа к данным через ВО инфраструктуру, а провайдеры данных применять сервисы для опубликования данных в ВО. Стандартами DAL обеспечивается схема, по которой центры данных, обсерватории и проекты смогут разработать ВО-совместимые сервисы. Для DAL разработаны три прототипа протоколов: ConeSearch, Simple Image Access (SIAP), Simple Spectrum Access (SSAP). ConeSearch возвращает данные из каталога или таблицы для указанного положения и радиуса поиска. SIAP возвращает указатели для изображений области неба, полученные по координатному запросу, а SSAP возвращает указатели для спектров и временных серий. Ведется работа по расширению DAL для других типов данных и встраиванию интерфейсов в существующие программные системы.

На примере SIAP [15] видно, что пока только рассмотрено решение технических проблем передачи данных и удовлетворение позиционного запроса с возможностью составления мозаик. Параметры, передаваемые в SIAP запросе, не включают характеристик качества данных. Они

отражают характеристику сервиса и запроса. Данные считаются однородными и идеально подготовленными для совместного сравнения и использования.

Протокол доступа к каталогам SkyNodes Interface обычно используется для запросов к нескольким каталогам, в отличие от SIAP и SSAP. Для удовлетворения запросов по этому протоколу используется расширенное подмножество SQL, называемое Astronomical Data Query Language (ADQL) [16]. ADQL, кроме координатных запросов, поддерживает доступ по протоколам ВО к таблицам, изображениям и спектрам. Многие астрономические каталоги хранятся в реляционных базах данных, поэтому SQL является пригодным средством для запросов. ADQL, в основном, использует подмножество оператора SELECT, с дополнительными функциями, позволяющими определять геометрические типы данных и выполнять операции над ними. SkyNode сервисы, определяются и применяются как XML веб-сервисы, используют для запросов ADQL. В базовом наборе операций ADQL предусмотрены запросы метаданных о грамматических спецификациях, таблицах, колонках, функциях, изображениях. С помощью этих операций мы можем получить информацию о

После такого краткого обзора деятельности рабочих групп IVOA можно сделать следующие выводы. Данные в инфраструктуре ВО пока считаются идеальными, и существующие спецификации IVOA предназначены для работы с science-ready данными. В разработке моделей данных, семантического словаря IVOA закладываются основы для представления и описания ошибок измерения физических величин, характеристик телескопов и астрономических приборов, которые определяют качество получаемых данных. В ВО регистрах в метаданные, описывающие ресурс, предполагается занесение характеристики качества ресурса, которая включает ошибки измерений, подтверждение корректности ресурса администратором регистра и общее описание. Идет опробование и отработка механизмов доступа к распределенным источникам, правда, пока без учета качества данных.

4 Научные данных

Астрономическую информацию можно разделить на четыре большие группы – наблюдения или «сырые», необработанные данные, готовые к научному анализу данные (science-ready data), теоретические модели и публикации. Рассмотрение всей совокупности астрономических данных в едином контексте с позиции качества является необходимостью и потребностью при интеграции информации в инфраструктуру ВО. У каждой группы данных свои критерии качества. Не будем рассматривать в сообщении теоретические модели и публикации а, обратимся к научным данным или иначе, обработанным, и астрономическим наблюдениям.

Производители наблюдательных данных, создатели каталогов или одним словом, – провайдеры информации – это астрономические организации и отдельные ученые. Градация ресурсов по объему и количеству огромная, от одной таблицы с десятком записей до нескольких тысяч каталогов с сотнями миллионов объектов. Имеются специализированные центры астрономических данных, которые обладают технической базой, позволяющей одновременное обращение по сети к ресурсам сотен и тысяч пользователей. В них работают специалисты, обеспечивающие сервисы доступа, хранение и курирование данных Примером такого центра может служить CDS. В нем собрана большая коллекция астрономических каталогов (~5400), которая постоянно пополняется из астрономических изданий. Есть ли в версии каталога ошибки, которые могут появиться при копировании? Маленькие таблицы можно проверить вручную, но каталоги с сотнями миллионов объектов сложно подвергнуть такой проверке. Как организовать такую верификацию и синхронизацию версий? Где отметить, что такая верификация и синхронизация выполнена, и иметь возможность просмотреть эту информацию? Она должна присутствовать в описании ресурса. В метаданные ресурса в регистрах ВО информация об идентичности версий с первоисточников не заносится и не контролируется.

Пользователю ВО необходима подробная информация о достоверности значений физических величин в данных, чтобы можно было анализировать приведенные к одной шкале величины. Он не должен беспокоиться о состоянии цифровых данных, правильность цифровых данных должна обеспечиваться инфраструктурой. Отдельные области каталогов или обзоров, которые астроном копирует из общего хранилища на рабочий компьютер, также должны содержать параметры качества, которые наследуются из основного

массива.

Исследователь, занимающийся сравнительным анализом каталогов, получает погрешности астрометрических и фотометрических привязок и может с уверенностью оценить однородность и качество этих ресурсов. Такая информация обычно публикуется в статьях. Ее можно было бы использовать для обратной связи между создателями ресурса и пользователями для улучшения качества данных и исправления промахов и пропусков в данных. Оценки качества данных каталога или обзора пользователями можно выставлять в описании ресурса в регистре, для того, чтобы исследователь или сервис мог принять решение о его использовании. Реализация обратной связи способствовала бы процессу улучшения качества данных в ВО.

5 Наблюдательные данные

Перейдем к проблемам качества наблюдательных данных. На результат наблюдения влияют много факторов: погодные условия (качество изображения, прозрачность), состояние аппаратуры (светоприемников, приборов), состояние телескопа, человеческий фактор. Наблюдение записывается в файл, обычно в FITS формате. FITS – самодокументируемый формат, позволяющий записывать в файл, кроме собственно наблюдательных данных, еще и параметры, описывающие процесс наблюдений и идентифицирующие данные. Разработка формата и принятие его в качестве астрономического стандарта для хранения и обмена данными было большим достижением в астрономии. Чем лучше описаны параметры наблюдение, тем лучше редукция, а, следовательно, надежнее и точнее результат.

При всех своих положительных качествах FITS стандарт имеет недостатки. Они связаны с тем, что стандарт был разработан достаточно давно – в начале 80-х прошлого века. И начальное его предназначение – перенос данных между различными операционными системами. Строго зафиксированы в стандарте были ключевые слова, определяющие представление данных. В начале века были переопределены и расширены ключевые слова для описания координатной системы изображений. Для остальных параметров нет жестких требований на наличие в заголовке файла. Параметры качества можно включать по желанию. Набор их не определен. В FITS стандарте не определены правила генерации новых ключевых слов. За время использования формата в разных обсерваториях появились свои клоны ключевых слов. Контроля над введением новых ключевых слов не велось. В результате для одного параметра наблюдения может использоваться несколько ключевых слов. В общем архиве САО РАН такие примеры имеются. Стандартизация ключевых слов для FITS формата ведется. Но, скорее всего с этой проблемой можно справиться с помощью UCDs. В CDS с проблемой множественности именований столкнулись раньше других, - разные названия колонок таблиц для одной величины. UCDs контролируется IVOA, имеются веб-сервисы для перевода, объяснения, проверки и назначения дескрипторов. В UCD есть дескрипторы для описания качества изображения, разного рода погрешностей.

Значения параметров наблюдения попадают в заголовок файла из разных 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, меньше пропусков и ошибок. Астрономическое программное обеспечение должно помогать детектировать ошибки телескопа, аппаратуры.

В ведущих обсерваториях мира достаточно давно используется понятие наблюдательного цикла, как единого технологического процесса. И в понятие наблюдательного цикла включено: подача заявок на наблюдательное время, составление расписания, подготовка к наблюдениям, сам процесс наблюдений, архивирование сырых данных, подготовка и проверка калибровочного материала, проверка правильности 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 разработкой и применением систем для обеспечения наблюдательного цикла. Такие системы требуют достаточных вложений и трудовых затрат, поэтому сейчас не везде используются. Необходимы стандарты и программное обеспечение для отдельных частей наблюдательного цикла. Разработка и применение таких стандартов – еще один шаг в сторону повышения качества данных.

Обычно обработка наблюдательных данных связана с последовательностью операций. Операции с изображениями – это арифметические действия, разного рода фильтрация, статистика. Для различных приборов последовательности операций при редукции данных могут отличаться. Последовательность обработки можно формализовать, контролируя качество данных и обработки с помощью фиксированных правил.

Желательно, чтобы эти операции выполнялись без участия человека в процессе формирования ответа на запрос, как некоторая последовательность операций, выполняемых одна за другой. Такие действия называются потоками или потоками работ/задач (workflow). В IVOA разрабатывается программное обеспечение для обработки информации, представляющее собой связанный набор распределенных сервисов с потоковым принципом выполнения. Потоковая парадигма для работы с распределенными данными состоит в повторном использовании простых сервисов для построения сложных приложений. Причем эти компоненты изолированы друг от друга посредством хорошо определенных протоколов, а именно определены правила для запуска, а также структура входных и выходных данных. В проекте AstroGrid (http://www.astrogrid.org) потоки работ в Common Execution Environment (CEA) выполняются для научных данных. И этот механизм может послужить прототипом для on-fly обработки наблюдений при обращениях к архивным данным в инфраструктуре ВО. Для этого необходима стандартизация последовательностей обработки (pipeline) для разных инструментов, создания библиотек таких процедур, разработка приемов автоматического контроля калибровочных данных. Поскольку архивы наблюдательных данных обсерваторий мира открыты для свободного пользования, то наблюдения не только крупных обсерваторий, но и малых должны стать составной частью ВО.

6 Хранение и сохранность данных

Из-за переменного характера некоторых астрофизических явлений долговременное хранение наблюдений обычно входит в компетенцию обсерваторий. Появление цифровых приемников излучения и записывающих устройств обеспечило астрономов большим числом наблюдений и компактными средствами хранения, которые уже невозможно рассматривать человеческим глазом, как фотографические пластинки. Цифровые носители требуют специального оборудования для раскодирования их содержимого. При считывании, записи, хранении могут возникнуть ошибки, потеря информации. Конечно, надежность устройств растет, и ошибки, возникающие при чтении и дублировании носителей информации, можно контролировать программно. Это нужно делать обязательно. 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, которые не гарантируют долговременной сохранности. Поэтому для поддержания в сохранности данных архива необходимо пересматривать технологию хранения каждые несколько лет и переносить содержимое архива с устаревших носителей на самые новые по технологии. Такую операцию приходится производить почти каждые 3-5 лет [17]. Качество данных связано с сохранностью информации на носителе, и важно, чтобы маркировка сохранности входила в некий общий сертификат качества, даже для одного файла.

7 Заключение

Мы попытались провести обзор факторов, влияющих на качество астрономических данных в контексте ВО. Интенсивная работа ведется в рабочих группах IVOA по обеспечению доступа к распределенным данным, и получены значимые результаты. Программные разработки по внедрению технологий ВО выполняются несколькими проектами, - NVO (USA), AstroGrid (UK), VOTech (EC), на базе крупных центров данных. Пока в запросах рассматриваются только science-ready данные. Созданием контролируемого IVOA семантического словаря UCDs сделан еще один шаг вперед в технологии семантического описания астрономических данных для обеспечения интероперабельности ВО систем. Основной упор делается на отладку технологии веб-сервисов, и деятельность групп подчинена этому, поэтому в спецификациях протоколов уделяется мало внимания описанию и контролю качества данных, параметры качества не участвуют в запросах. В разработке доменной модели данных в астрономии

этому понятию уделено должное внимание. Из-за сложности обработки сырых данных включение архивов наблюдательных данных в ВО не являются близкими планами Альянса. Быстрорастущие объемы данных и сложности их анализа усиливают роль систем управления базами данных в астрономии. Хранение цифровой информации в быстро меняющемся мире цифровых носителей и компьютерных систем остается дорогой по материальным затратам и трудно решаемой задачей.

Литература

1] Can Astronomy Manage Its Data?, Norris R., IAU Bulletin 96

2] Report to the 24th CODATA General Assembly on data activities in the International Astronomical Union, Ed. by Norris R., 2004. http://www.atnf.csiro.au/people/rnorris/WGAD/IAUCodataReport2004.htm

3] Special Session 6 at the 26th IAU General Assembly, Astronomical Data Management, http://www.astronomy2006.com/special-sessions.php#sps6

4] Public Access to Astronomical Archives. The Resolution of 5 Commission of IAU adopted by the IAU General Assembly on 24 July 2003. http://www.atnf.csiro.au/people/rnorris/WGAD/Resolution.htm

5] 2020 Computing: Science in an exponential world, Szalay A.&Gray J., http://www.nature.com/news/2006/060320/full/440413a.html

6] The UCD1+ controlled vocabulary, Version 1.11, IVOA Recommendation 31/12/2005, http://www.ivoa.net/Documents/latest/UCDlist.html

7] Resource Metadata for the Virtual Observatory, Version 1.01, IVOA Recommendation 2004/04/26, http://www.ivoa.net/Documents/REC/ResMetadata/RM-20040426.html

8] CDS GLU http://simbad.u-strasbg.fr/glu/cgi-bin/gludichelp.pl?frame=all

9] IVOA Identifiers, V1.10, IVOA Proposed Recommendation 03 March 2005, http://www.ivoa.net/Documents/latest/IDs.html

10] Virtual Observatory Data and Service Locator, http://nvo.stsci.edu/VORegistry/index.aspx

11] A unified domain model for astronomy, for use in the Virtual Observatory, Version 0.9, IVOA Working Draft 2003-11-04, http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel/DomainModelv0.9.1.doc

12] Data Model for Observation, V.0.23, IVOA DM WG Internal Draft, 2004-05-16, http://www.ivoa.net/internal/IVOA/IvoaDataModel/obs23.pdf

13] Data Model for Quantity, V.0.23, IVOA DM WG Internal Draft, 2004-05-17, http://www.ivoa.net/internal/IVOA/IvoaDataModel/qty23.pdf

14] Charter for the Semantics WG, WD 2006/05/17 http://www.ivoa.net/internal/IVOA/IvoaUCD/Charter-Semantics.pdf

15] Simple Image Access Specification Version 1.00, IVOA Working Draft 24 May 2004, http://www.ivoa.net/Documents/latest/SIA.html

16] IVOA Astronomical Data Query Language, Version 1.051, IVOA Working Draft 13 Jul 2006, http://www.ivoa.net/Documents/WD/ADQL/ADQL#20060712.doc

17] Using DVD technology for archiving astronomical data, Pirenne B., Albrecht M., http://archive.eso.org/~archeso/DVD_in_astronomy.html

Astronomical data quality at a context of resource integration into the virtual observatory

Zhelenkova O.P., Vitkovsky V.V., Plyaskina T.A.

A virtual observatory (VO) is a web infrastructure to unify multiply digital astronomical collections. The science ready data need for science analysis with VO facilities. Constantly growing volumes of information push the astronomical community to realization of a holistic approach to data quality providing.