Федеративные принципы построения интегрированного банка из разнородных коллекций знаний

© Желенкова О.П., Витковский В.В., Калинина Н.А., Шергин В.С., Черненков В.С.

Специальная астрофизическая овсерватория РАН
zhe@sao.ru

Аннотация

Изучение физики небесных объектов невозможно без сравнительного изучения наблюдательной информации, полученной в разных диапазонах электромагнитного спектра. Многие астрофизические исследования основываются на использовании уже имеющихся в наблюдательных архивах и цифровых обзорах данных, объем которых в последнее десятилетие растет по экспоненте. Актуальной является задача создания специальной инфраструктуры, объединяющей эти информационные ресурсы и использующей Интернет для проведения исследований. В Специальной астрофизической обсерватории РАН (САО РАН) разрабатывается интегрированный банк данных (ИБД). На примере архива наблюдательных данных обсерватории, который является частью ИБД, рассматриваются принципы федеративного объединения разнородных данных и доступа к ним в соответствии со стандартами и спецификациями International Virtual Observatory Alliance (IVOA)[1].

1 Концептуальные основы виртуальной обсерватории.

Еще несколько лет назад изучение Вселенной шло от создания и развития приборной базы при овладении новым диапазоном электромагнитного излучения. Очевидно, что открытие новых объектов, изучение физики небесных объектов невозможно без сравнительного изучения наблюдательной информации, полученной в разных диапазонах электромагнитного спектра. Решение этой проблемы должно в первую очередь основываться на объединении уже имеющихся данных, так в астрономии имеются цифровые обзоры неба: 2MASS, SDSS, POSS-2, FIRST, COBE, MAP, ROSAT, GALEX, которые покрывают все небо по 15 различным частотным диапазонам электромагнитного излучения, - от рентгеновского, ультрафиолетового, оптического, инфракрасного, микроволнового до радио[1].

Объединению этой информации препятствует используемая в наблюдениях методическая база, которая осталась прежней. Информация, полученная в разных частотных диапазонах, имеет разные характеристики: угловое и частотное разрешение, координатное обеспечение, чувствительность, форматы регистрируемых и хранимых данных. Понятно также, что при нынешнем состоянии компьютерной техники и телекоммуникаций выполняются единичные исследовательские работы для небольшого числа объектов, но они не могут обеспечить существенного прогресса в изучении Вселенной.

Произвести поиск новых знаний в уже имеющейся и открытой информации цифровых обзоров неба и архивов наблюдений можно только с реализацией массовых запросов. Актуальной является задача создания специальной инфраструктуры, формирование среды, включающей архивы, каталоги данных, программное обеспечение и использующей Интернет для проведения научных исследований[2].

При этом понятно, что необходимо решить следующие проблемы[3]:

- преодолеть разнородность форматов данных;
- реализация координатного поиска и проблема калибровок данных, приведение к единой шкале по потокам;
- организация сетевого сервиса с учетом того, что базы данных могут быть большого объема, территориально разнесены.

Совместное решение этих проблем, выбор и создание стандартов послужит базисом для интероперабельного объединения разнородных цифровых коллекций (обзоры, каталоги, наблюдательные архивы), что является одной из концептуальных основ виртуальной обсерватории[1,4].

2 Информационные ресурсы обсерватории

САО РАН – обсерватория, обеспечивающая работу крупнейших российских телескопов - БТА и

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

РАТАН-600. Оба телескопа являются сложными экспериментальными комплексами, наблюдения на которых выполняются как российскими, так и иностранными исследователями. Оба телескопа – принципиально многопрограммные и многопользовательские системы и остаются до сих пор рекордными по количеству получаемой научной информации. За 30 лет работы обсерватории накоплен уникальный наблюдательный материал в оптическом и радио диапазонах.

В САО РАН систематические работы по архивизации наблюдательных данных начались с 1985 года [5]. На основе этих работ была предложена концепция Банка данных обсерватории как объединенного хранилища информации. За прошедшие годы увеличилось как число коллекций цифровых данных, так и их объем. Это разнообразные каталоги астрономических объектов, цифровой обзор неба, web-ресурсы, созданные в обсерватории. Эти коллекции активно используются. Соответственно этому актуальность задачи формирования и пополнения Банка данных обсерватории, а также обеспечения широкого и удобного доступа к данным с развитием техники и технологии наблюдений только возрастает. Развитие этого подхода состоит в объединении информационных ресурсов обсерватории и создание интегрированного Банка данных (ИБД) обсерватории.

1.1 Архив наблюдательных данных

Архив наблюдательных данных является одной из основных составляющих ИБД. Многие практические решения задач хранения и работы с данными архива легли в основу организации банка данных. Архив за время своего существования прошел несколько этапов развития в смысле представления и хранения данных, предоставления сервисных функций. Эти этапы тесно связаны с развитием компьютерных и информационных технологий, а также с необходимостью интеграции с астрономическим сообществом в проведении научных исследований. Можно рассмотреть этапы развития архива наблюдательных данных обсерватории в виде следующей цепочки:

локальный архив -> основной архив -> общий архив.

Приведем более подробную характеристику каждого этапа. В *локальном архиве* хранятся данные, производимые одним методом наблюдений или несколькими сходными. Каждый метод наблюдения связан с определенным компьютерно-аппаратным комплексом - системой сбора. Выход системы сбора - это архив файлов определенного формата со своими параметрами, структурой и размерностью массивов. В архив погружаются текущие наблюдательные данные, обеспечивается сохранность данных и поддержка одного типа формата данных.

Основной архив - коллекция локальных архивов. Архив является прозрачным для пользователя, то есть, он не меняет форматы и параметры хранящихся данных; в каком формате поступили данные на вход архива, в таком их и получили при запросе. Для данных используется FITS-формат/FITS-подобный формат.

Общий архив - основной архив с организацией сетевого доступа/web-доступа.

Следствием многопрограммности телескопов является существование и разработка отличающихся между собой по многим параметрам систем сбора данных. Одной из существенных основных проблем при формировании общего архива обсерватории является различие параметров выходных данных и их форматов. Сложилось так, что форматы цифровых данных, регистрируемых системой сбора, разрабатывались и реализовывались научно-исследовательской группой, занимающейся созданием прибора, поэтому в наблюдательных файлах разных методов наблюдения нет единого формата данных и одного множества ключевых слов, используемых для описания этих файлов. Поскольку системы сбора модернизируются, то меняются форматы данных и параметры, описывающие наблюдение, поэтому каждый архив включает в себя не одну версию форматов данных. Еще одной особенностью локальных архивов является существенное разнообразие данных: многочастотность, методы получения и обработки и структурная разнородность.

Так например, общий архив обсерватории на текущий момент состоит из 14 локальных архивов. Число параметров, используемых для описания наблюдения, варьируется от десятка до двух сотен параметров в зависимости от типа локального архива (типа прибора). Наблюдение, как семантическая единица архива, в большинстве случаев сохраняется в одном файле, но может сохраняться и в нескольких. Имеются tar-архивы, которые содержат одно наблюдение, но данные этого наблюдения хранится в нескольких файлах – двоичное изображение, текстовый файл с параметрами и текстовый файл со служебной информацией. В некоторых локальных архивах несколько наблюдений записываются в один tar-архив.

3. Федеративные принципы функционирования ИБД

Для реализации архива используются следующие базовые принципы [6], которые суммируют опыт работы с разнородными локальными архивами:

- основной смысловой единицей архива является наблюдение;
- архив является прозрачным для пользователя, то есть, он не меняет форматы и параметры хранящихся данных; в каком формате

поступили данные на вход архива, в таком их и получили при запросе;

- в архив погружаются текущие наблюдения; наблюдательные данные могут копироваться в архивах пользователей по их запросам с любого архивного уровня;
- "старые" наблюдения подгружаются в архивную систему по требованию;
- исключительное авторское право использования данных архива, содержащих информацию об астрофизических объектах, в течение 2 лет после выполнения наблюдений принадлежит заявителям наблюдательной программы;
- не проводится строгой границы между первичными и обработанными данными, то есть, возможно хранение частично-обработанных и обработанных данных.

Продолжая и развивая эти принципы интегрированный банк не накладывает ограничений на внутреннее функционирование и пополнение локальных архивов и берет на себя выполнение общих функций – взаимодействие с внешним пользователем, выполнение внешних правил взаимодействия и защиты. В сущности, реализуется принцип *Федерации:* выполнение внешних правил и сохранение внутренних структур, аналогично с принятым в государственной политике. Понятно, что федерация накладывает некоторые ограничения на субъекты, но эти ограничения минимальны.

Кроме экспериментальных данных интегрированный банк должен включать объекты другой природы, семантики и назначения, к примеру: каталоги, публикации, цифровые обзоры. Чтобы астрономическое сообщество имело возможность пользоваться информационными ресурсами обсерватории, необходимо реализовать web-сервисы, соответствующие разрабатываемым стандартам.

Информационные сервисы являются внешними правилами поведения, и в настоящий момент астрономическим сообществом предлагается прототип технологии их выполнения. В него включены спецификации на регистрацию сервисов [7], тип запросов [8,9], формат ответа на запрос [10,11]. Для установления семантической связи между параметрами/ключевыми словами, описывающими каталог, изображение, таблицу, и их физическим значением используется внешний тезаурус [12].

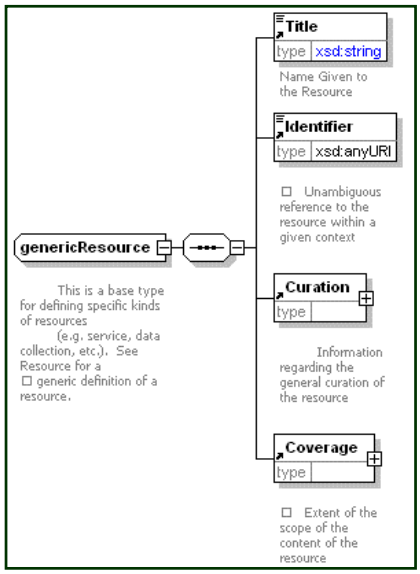
Внутренние правила функционирования ИБД опираются на вышеназванные принципы и, прежде всего, сохраняются внутренние форматы и правила пополнения коллекций, к которым относятся общий архив наблюдательных данных обсерватории, каталоги и обзоры, которые поступают в обсерваторию из других астрономических учреждений.

4. Спецификации и стандарты IVOA – правила внешнего взаимодействия

Рассмотрим подробнее принятые IVOA спецификации для демонстрации возможностей виртуальной обсерватории. Спецификации описывают простые сервисы, которые должны реализовываться информационными ресурсами ВО. Эти сервисы позволяют организовать запрос по выбору изображений интересующего пользователя участка неба из архивов, каталогов, цифровых обзоров, расположенных в различных обсерваториях. В их состав входят уже упомянутые в предыдущем параграфе: регистрация ресурса, координатный запрос и получение ответа в определенном формате. Эти спецификации предполагается использовать для реализации внешнего web-взаимодействия с архивом.

Правила регистрации и описания ресурса определяются в документе “Resource and Service Metadata for the Virtual Observatory” [7]. Для информационных служб астрономии, чтобы участвовать в ВО, предлагается иерархическая система для описания метаданных. На высшем уровне этой иерархии требуется минимальное количество информации, достаточной, прежде всего для того, чтобы обратить внимание на существование ресурса и описать его. На более низких уровнях метаданные более подробны, чтобы иметь возможность описывать синтаксис запроса, протоколы доступа, политику использования.

В этой спецификации вводятся следующие понятия: *resource* – элемент ВО (коллекция данных, инструмент,…, то, что поддерживается и имеет уникальный идентификатор); *organization* – (обсерватория, проект, университет или отдельный исследователь; – пользователь может обратиться к любому элементу ВО для выполнения каких-то действий; query service – поддерживает протокол запрсов/ответов; *registry* - служба запроса, для которого ответ является структурным описанием



других услуг, что позволяет пользователю указывать спектр сервисов, которые имеются у ресурса.

Ресурс – это коллекция одного или нескольких сервисов, нескольких ресурсов с одним описанием. Для описания ресурсов и сервисов используются метаданные. На рисунке 1 приведена диаграмма XML Schema документа для описания ВО ресурса.

Следующая спецификация описывает процедуру и интерфейс (Simple Image Access Interface – SIM) передачи изображений из различных коллекций астрономических данных [8,9]. Требования, которые должны выполняться при передачи изображений:

- web метод по запросу изображения позволяет пользователям искать изображения выбранного участка неба. Параметры передаются как HTTP GET запрос, причем необходимо обеспечить передачу основных параметров: координат центра выбранной области и ее размер и еще необязательных дополнительных параметров, описывающих геометрию изображения, формат и т.д.;
- передача полученного изображения выполняется в формате VOTable [10] с набором обязательных элементов этого формата, которые описывают передаваемое изображение с указанием ресурса, координат, статуса выполнения запроса, формата файла. Названия параметров должны соответствовать домену имен VOX:namespace, который определяется по тезаурусу UCD[12]. Успешное завершение передачи и ошибки, если таковые имеются, подтверждаются соответствующими сообщениями.

Сервисы изображений делятся на четыре категории:

- *Image Cutout Service*. Передача прямоугольной площадки без выполнения масштабирования изображения.
- *Image Mosaicing Service*. То же самое, что и предыдущее, но добавляется возможность изменения размера, масштаба и проекции изображения.
- *Atlas Image Archive*. Получение специально предвычисленной области, для работы с большими участками неба.
- *Pointed Image Archive*. Доступ к коллекции маленьких участков неба.

5. Реализация принципов ИБД на примере Общего архива наблюдательных данных обсерватории

Механизм перевода внутренних представлений цифровых данных во внешние при выполнении web-сервисов рассматривается на примере локальных архивов, которые входят в общий архив наблюдательных данных.[1] По опросам пользователей определены несколько типов запросов к архивным данным. К ним относятся запросы по: дате наблюдения, прибору, типам файлов, координатам, имени источника, программе наблюдений, автору и наблюдателям. Анализ источников информации (с точки зрения выполнения запроса) проводился по следующей схеме:

- достаточно ли ключевых слов в заголовке для выполнения запроса,
- необходимо ли привлечение других источников, кроме заголовка наблюдательного файла,
- какие именно источники необходимы,
- правила (связь между ключевыми словами), по которым возможен поиск соответствий запросу.

В локальных архивах данные имеют разные наборы параметров для описания внутреннего представления и физических параметров, поэтому был проведен анализ того, какой тип внешних запросов может удовлетворять цифровая коллекция.

Поиск по дате наблюдения может выполняться по всем локальным архивам [13]. Реализованы сервисные функции, предоставляемые общим архивом пользователю:

- пользовательский интерфейс с использованием web-браузеров;
- организация выдачи выбранных данных по сети;
- запросы по дате, методу и дате,
- просмотр FITS-заголовка,
- предварительный просмотр данных для файлов в FITS-формате путем генерации черно-белой катринки в jpg-формате,
- получение данных наблюдательного сета из архива расписаний (запрос по названию программы наблюдений).

Возможность координатного запроса, который описан в спецификации SIM, применим к части коллекций, поскольку не в каждом локальном архиве данные имеют набор координатных параметров. Для локальных архивов, в которых хранятся прямые снимки участков неба можно опробовать сервис изображений Pointed Image Archive. В SIM не описан сервис для передачи спектральных данных, когда не определяются размеры площадки.

6. Заключение

При подготовке и включении цифровой коллекции в интегрированный банк необходимо следовать некоторой процедуре: иметь полное семантическое описание параметров и форматов данных, определить типы запросов, контролировать именование и формат ключевых слов по локальному тезаурусу. В нашем случае каждый локальный архив является ресурсом, для которого необходимо составить описание, соответствующее спецификации “Resource and Service Metadata for the Virtual Observatory”.

Именование и формат параметров, описывающих наблюдение не всегда соответствуют стандарту [11]. Имеются варианты именований для параметров с одним физическим смыслом. Поэтому необходимо создание промежуточного (локального) тезауруса для поиска соответствия внутренних ключевых слов внешним UCD-названиям физических параметров наблюдений[12].

Разработка и тестирование федеративных принципов работы с информационными ресурсами на примере ИБД САО РАН направлено на реализацию следующих принципов построения Российской виртуальной обсерватории (РВО):

- объединение информационных архивов и информационных источников в одну распределенную систему;
- связывание в такой системе со стандартизованными интерфейсами и средствами Интернет доступа архивов и центров данных, экспериментальных комплексов РВО.

Литература

[1] Quinn P. J., The Astrophysical Virtual Observatory. Drivers, Status and Planning, in conf. “Toward an International Virtual Observatory”, Garching, 8-14 June, 2002 (in press).

[2] Szalay, A.S., Brunner, R.J., Astronomical Archives of the Future: a Virtual Observatory. astro-ph/9812335 (1998)

[3] R. Williams, Approaches to Federation of Astronomical Data. Virtual Observatories of the Future, ASP Conference Series, v.3, 2000, pp1-15

[4] Szalay, A.S., The National Virtual Observatory, Astronomical Data Analysis Software and System X, ASP Conf. Ser., v.238, pp 3-12, 2001

[5] Кононов В.К., Моносов М.Л., Витковский В.В., Липовецкий В.А., Архив наблюдательных данных САО АН СССР. Принципы организации. Сообщения САО, 65,32-44, 1990

[6] V. Vitkovskij et al., The project of distributed information system OASIS, Baltic Astronomy, v.9, №4, pp 578-582, 2000

[7] Resource and Service Metadata for Virtual Observatory. Version 0.7. IVOA Working Draft, 2003, http://www.ivoa.net/

[8] Simple Image Access Prototype Specification, 2002, http://www.us-vo.org/news/simspec.html

[9] Conesearch, 2002, http://www.us-vo.org/metadata/conesearch/index.html

[10] VOTable: A Proposed XML Format for Astronomical Tables, R. Williams et al., 2001, http://cdsweb.u-strasbg.fr/doc/VOTable

[11] Definition of the Flexible Image Transport System (FITS), R.J. Hanish et al., A&A, v.376, pp 359-380, 1986

[12] Unified Content Descriptors (UCDs), 1999, http://vizier.u-strasbg.fr/doc/UCD.htx

[13] Общий архив наблюдательных данных САО РАН, 2002, http://www.sao.ru/oasis/cgi-bin/fetchru

Federal principles of the integrated bank creation from heterogeneous knowledge collections

Zhelenkova O.P., Vitkovskij V.V., Kalinina N.A., Shergin V.S., Chernenkov V.N.

The investigation of physics of celestial objects is impossible without comparative learning of the observation information obtained in different ranges of an electromagnetic spectrum. Many astrophysical researches are grounded on usage of the data already available in observation archives and digital surveys, which size per the last decade grows exponentially. The task of creation of a special infrastructure joining these information resources and utilizing the Internet for science researches is actual. The integrated data bank (IDB) is developed in the Special astrophysical observatory of RAS (SAO RAS). On an example of the observation data archive, which is a part of the IBD, the federal association principles of the heterogeneous data and access to them are considered according to standards and specifications International Virtual Observatory Alliance (IVOA) [1].

[1] Работа выполняется при поддержке РФФИ, грант № 07-03-90034