ПРИМЕНЕНИЕ XML К ОПИСАНИЮ РАЗНОРОДНЫХ НАБЛЮДАТЕЛЬНЫХ ДАННЫХ

Витковский В.В., Желенкова О.П., Калинина Н.А., Шергин В.С.,
Черненков В.Н., Малькова Г.А.
Специальная астрофизическая обсерватория РАН, п. Нижний Архыз,
369167, Россия
e-mail: zhe@sao.ru

THE USE OF XML TO THE DESCRIPTION OF DEVIRSE OBSERVATIONAL DATA

Vitkovskij V.V., Zhelenkova O.P., Kalinina N.A., Shergin. V.S.,
Chernenkov V.N., Mal'kova G.A.
Special Astrophysical Observatory of RAS, Nizhnij Arkhyz, 369167, Russia
e-mail: zhe@sao.ru

The Internet using for scientific researches and formation of specialized environment including interoperable archives, catalogues, software is an urgent task in world astronomical community. «Virtual telescope», «virtual observatory» are uniting name for the innovation development on interoperability and scalability of on-line data.

The FITS-format is the most widespread way of astronomical data storage now, but it does not allow make effective search of the necessary information, especially in conditions of constant growth of data volume. Existing industrial information technologies of management and performance of the data, such as database management systems, web-access, XML, Java, and their fast development give hope for realization of information resources interoperability in the close future. The virtual observatory concept forces in a new fashion the consideration of questions got in touch with science archives.

SAO RAS archive consists from local archives that to receive observational data from diverse acquisition systems and the one is a good test example for development of similar systems. We have put a task to analyze the available data and to organize archival system for transparent access to diverse local archives, giving attention to data quality, automation of data re-processing and integration with the astronomical catalogues and surveys.

В последние десятилетия из-за быстрого развития телекоммуникаций, компьютерных и информационных технологий в астрономии происходят коренные изменения в методах проведения научных исследований. Современные наземные и спутниковые телескопы предоставляют возможность наблюдений в полном диапазоне электромагнитного спектра. Новое поколение больших телескопов, автоматизированные системы сбора и ис-

пользование панорамных цифровых приемников электромагнитного излучения повысили эффективность наблюдений и увеличили на порядок объемы получаемой информации. Рост вычислительных мощностей компьютерного оборудования и развитие программного обеспечения позволяют накапливать, обрабатывать и сохранять терабайтные объемы данных. Новые астрономические проекты требуют международной координации исследований, включают географически разнесенные ресурсы, данные и команды участников.

Актуальной является задача создания специальной инфраструктуры, формирование среды, включающей архивы, каталоги данных, программное обеспечение и использующей Интернет для проведения научных исследований. «Виртуальный телескоп», «виртуальная обсерватория» - объединяющее название для этих инновационных разработок. С 2000 года ведется в США проект “National Virtual Observatory”[1], немного позднее начались работы в Европе – “Astrophysical Virtual Observatory”[2], “AstroGrid”[3]. Аналогичные работы проводятся в России – «Российская виртуальная обсерватория» (САО РАН, ИНАСАН) [4].

В июне 2002 года на международной конференции “Toward an International Virtual Observatory”(Гаршинг, Германия) создан международный консорциум IVOA (International Virtual Observatory Alliance), в него вошли национальные проекты. На конференции рассматривались возникающие при реализации «виртуальной обсерватории» проблемы и была выработана следующая стратегия для их решения [5]:

Стратегия VO (Virtual Observatory)

Проблема	Подходы к решению	
Медленный рост скорости вычислений	Распределенные вычисления	GRID
Ограниченный объем хранилищ данных	Распределенные данные	
Ограниченная пропускная способность каналов связи	Иерархическая структура информации (перемещать только то, что необходимо пользователю)	
Разнородность данных	Интероперабельность	Стандарты VO

Объем данных, получаемых в наблюдательных экспериментах, ежегодно удваивается. В качестве примера можно привести данные об архиве крупнейшей европейской обсерватории ESO (European Southern Observatory) [5]:

Архив наблюдательных данных ESO

	1998	2002
Входной поток данных	100 GB/year	1100 GB/year
Общий объем данных	0.5 TB	10 TB
Объем запрошенных данных	100 GB	2-3 TB

Архивы, содержащие наборы данных, в большинстве своем в форме когерентных обзоров, часто в нескольких диапазонах, причем больших участков неба, открывают новые возможности для добычи знаний из имеющихся данных. Следует отметить, что астрономические данные не теряют своей научной значимости с течением времени.

Интероперабельность - организация обмена информацией между разнородными базами данных и информационными службами как распределенного запроса, прозрачного (seamless) для пользователя, в смысле доступа к интересующим данным. Этот сервис предоставит уникальные возможности для выполнения научных исследований, как-то: поиск новых объектов, мониторинг переменности, изучение физических свойств известных объектов с привлечением имеющейся информации в разных диапазонах электромагнитного спектра.

Обработка запроса [6] к необходимым информационным ресурсам обычно может потребовать нескольких шагов:

1. Поиск ресурса: какие источники могут обеспечить необходимой информацией;
2. Описание ресурса: адреса источников и синтаксис запроса;
3. Выполнение запроса;
4. Представление информации о полученных данных (к примеру, объем данных);
5. Получение данных (со сравнением и кросс-корреляцией);
6. Визуализация данных.

Поиск ресурса может выполняться с помощью web-сервисов, типа промышленных стандартов UDDI (Universal Description, Discovery and Integration) или GLU (Generateur de Liens Uniformes)[7] – астрономического регистра сервисов. Для выполнения запроса предполагается использовать новые протоколы для передачи информации – WSDL, SOAP, для представления данных - новые форматы (для астрономии) – XML.

Специальная астрофизическая обсерватория (САО РАН) - обладатель двух крупнейших телескопов России, имеющих статус инструментов коллективного пользования. В обсерватории более 20 лет ведется цифровой архив наблюдательных данных. За это время накоплен уникальный наблюдательный материал в оптическом и радиодиапазоне. На радиотелескопе архивизация цифровых данных проводится с 1979, на БТА с 1988 с введением в штатный режим цифровых светоприемников.

На инструментах для получения астрофизических данных за этот срок использовались более 40 методов наблюдений. Каждый метод наблюдения связан с определенным компьютерно-аппаратным комплексом - системой сбора. Входом локального архива является выход системы сбора. Сложилось так, что архивизация данных велась научно-исследовательской

группой, разрабатывающей прибор для наблюдений, поэтому в архиве нет единого формата данных

В начале 80-х астрономы стали широко использовать FITS-формат (Flexible Image Transport System)[8] для обмена данными. Это самодокументируемый формат, содержащий описание данных и сами данные в одном файле. В САО FITS-формат впервые (1988)[9] стал использоваться для данных ПЗС (Прибор с Зарядовой Связью)-камеры. Основная часть данных архива сохраняются в FITS или в форматах, близких к FITS по структуре.

В том виде, в котором он существует сейчас, архив поддерживается с 1995 года. В него входят необработанные (raw) и калибровочные данные, получаемые на телескопах обсерватории. Для хранения используются оптические диски. Имеются данные на ленточных носителях. Ведутся работы по восстановлению и перезаписи "старых" данных в существующие форматы.

Архив наблюдательных данных САО РАН (CD-диски)

	1998	2001
Общий объем данных	57 GB	98 GB
Входной поток данных	10 GB/year	
Среднесуточный поток данных: • оптика • радио диапазон	25 MB/day 5 MB/day	

Архив САО за 20 лет 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 мировым астрономическим сообществом в проведении научных исследований.

В Отделе информатики обсерватории с 1999 года разрабатывается проект архивной системы обсерватории [10]. В состав архивной системы входят: каскадная система архивизации, информационно-поисковая система с использованием СУБД и архив наблюдательных данных. К основным принципам архивной системы относятся:

- в архиве предполагается хранить результаты всех наблюдений, выполненных на телескопах обсерватории;
- основной смысловой единицей является наблюдение;
- не меняются 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 уровня каскадной системы архивизации; "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.

Основные требования к архивной системе:

- надежное хранение данных;
- сетевой доступ к данным;
- контроль доступа;
- отсутствие жестких ограничений на форматы хранимых данных.

Сервисные функции архивной системы:

- удобный пользовательский интерфейс с использованием web-браузеров;
- архивизация необработанных данных с предоставлением необходимых данных для выполнения предварительной обработки;
- стандартные и параметрические запросы;
- организация запросов по нескольким локальным архивам;
- предварительный просмотр отобранных данных.

Передаваемые из систем сбора данные должны иметь полное семантическое описание, то есть, физический смысл параметров и данных должен быть известен. Описание должно содержать, как минимум, информацию для однозначной идентификации наблюдения и его обработки. Каждая система сбора имеет свой стандартный и зафиксированный формат выходных данных. Предполагаемые изменения в выходных форматах предварительно согласовываются. Наблюдение сохраняется в архиве в том виде, в котором оно было передано из системы сбора.

Для погружения данных в архив используется каскадная схема архивизации, которая имеет четыре уровня хранения:

- первый уровень - буферный архив;
- второй уровень - промежуточный архив;
- третий уровень - оперативный архив;
- четвертый уровень - CD-библиотека.

Данные переходят последовательно с одного уровня хранения на другой. Буферный архив - это область дискового пространства выделенного файл-сервера. По окончании наблюдательной программы в этой области хранятся все полученные данные, затем они переносятся в промежуточный архив на специализированном сервере, где производится подготовка образа диска и запись на CD-диски. Постоянное хранение данных – CD-библиотека. В оперативном архиве производится просмотр, коррекция параметров файлов с наблюдениями. Данные в оперативный архив подгружаются по

запросу. Оперативный архив располагается на архивном сервере, где также расположена информационно-поисковая система.

В связи с развитием и изменением базовых принципов астрономических наблюдательных архивов, расширением их сервисных функций развитие архивной системы предполагается выполнять так, чтобы соответствовать уровню подобных систем. Необходима организация интероперабельности локальных архивов.

Можно рассмотреть этапы развития архивной системы в виде следующей цепочки: ***локальный архив -> основной архив -> общий архив -> хранилище ->научный архив -> "VO-ready" архив.***

Первые два этапа реализованы и поддерживаются, третий этап разрабатывается, остальные – проектируются. Приведем более подробную характеристику каждого этапа.

Локальный архив - данные на машинных носителях, производимые одним методом наблюдений (несколькими сходными). В таком архиве хранятся результаты всех наблюдений, выполненных данным методом, в архив погружаются текущие наблюдательные данные, обеспечивается сохранность данных и поддержка одного типа формата данных. Доступ к данным осуществляется через ответственного за метод наблюдения.

Основной архив - коллекция локальных архивов. В архиве имеется порядка 15 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. Доступ к данным – через администратора архива. Реализован первый и второй уровень каскадной системы архивизации.

Общий архив - основной архив с организацией сетевого доступа/web-доступа к данным (локальная сеть, Интернет). На этом уровне добавляется информационно-поисковая система для преодоления гетерогенности данных локальных архивов.

В информационно-поисковой системе имеются три уровня хранения: SIB (Service Information Block), FITS-заголовок, наблюдение.

Первый уровень - основа для выполнения запросов. Здесь на каждое наблюдение имеется сервисный информационный блок (SIB), в котором хранятся параметры для организации стандартных запросов и авторизации доступа. Второй уровень информационно-поисковой системы - FITS-заголовки, которые можно просматривать. Второй уровень необходим, если нет возможности организовать on-line доступ сразу ко всем файлам. Третий уровень – хранение наблюдательные данные для организации передачи по сети, визуализации наблюдательных данных. Сервисные функции, предоставляемые общим архивом пользователю:

- пользовательский интерфейс с использованием web-браузеров;

- предоставление необходимых данных для выполнения предварительной обработки (калибровки);
- организация выдачи выбранных данных по сети;
- расширение типов запросов (по дате, методу и дате, имени источника, координатам, просмотр FITS-заголовка, по типу файла).

На этом этапе предполагается реализовать каскадную схему архивизации полностью, причем сведения о выполненном наблюдении сразу помещаются в информационно-поисковую систему с указанием уровня хранения.

Хранилище - "чистые" данные; в параметрах, описывающих каждый файл, исправлены возможные ошибки (дата, название программы, метеорологические параметры). Выполняется контроль значений параметров в SIB с использованием данных, хранящихся в базе данных по расписанию наблюдений и инструментальной базе данных. Осуществляется динамическое пополнение словаря ключевых слов после просмотра заголовков файлов.

Для доступа к хранимой информации предполагаются следующие сценарии:

- всем доступны данные, у которых закончился авторский период, и служебные данные;
- доступ с предварительной регистраций (контроль работы с архивными файлами);
- три уровня доступа: администратор, авторы программ, пользователь. Для авторизации доступа необходимо взаимодействие с базой данных расписания наблюдений (заявители программ, ответственные наблюдатели).

Научный архив - стандартизация процедур обработки и оценки качества калибровочных файлов, автоматическая стандартная обработка наблюдательных данных.

Виртуальный телескоп – организация архивной системы, как этого требует концепция виртуальной обсерватории. Это потребует интеграции с каталогами, подготовки архива для реализации распределенных GRID-вычислений. Необходимо обеспечить достаточно удобную процедуру добавления в архив новых типов локальных архивов.

Наиболее распространенный в настоящее время способ хранения астрономических данных, FITS-формат, не позволяет эффективно производить поиск необходимой информации. Консорциум IVOA предлагает использовать стандартом для описания данных формат VOtable 1.0 [11]. Предполагается использовать его для создания метаописаний локальных архивов. Этот формат - предлагаемый XML-формат для представления табличных данных в контексте виртуальной обсерватории.

```
<?xml version="1.0"?>
<!DOCTYPE VOTABLE SYSTEM "http://us-vo.org/xml/VOTable.dtd">
<VOTABLE version="1.0">
```

```
<DEFINITIONS>
        <COOSYS ID="myJ2000" equinox="2000." epoch="2000." system="eq_FK5"/>
</DEFINITIONS>
<RESOURCE>
        <PARAM name="Observer" datatype="char" arraysize="*" value="William
Herschel">
                <DESCRIPTION>This parameter is designed to store the observer's name
                </DESCRIPTION>
        </PARAM>
 <TABLE name="Stars"> <DESCRIPTION>Some bright stars</DESCRIPTION>
        <FIELD name="Star-Name" ucd="ID_MAIN" datatype="char" arraysize="10"/>
          <FIELD name="RA" ucd="POS_EQ_RA" ref="myJ2000" unit="deg"
        datatype="float"                precision="F3" width="7"/>
        <FIELD name="Dec" ucd="POS_EQ_DEC" ref="myJ2000" unit="deg"
        datatype="float" precision="F3" width="7"/>
        <FIELD name="Counts" ucd="NUMBER" datatype="int" arraysize="2x3x*"/>
        <DATA> <TABLEDATA>
                <TR> <TD>Procyon</TD><TD>114.827</TD><TD> 5.227</TD>
                <TD>4 5 3 4 3 2 1 2 3 3 5 6</TD> </TR>
                <TR> <TD>Vega</TD><TD>279.234</TD> <TD>38.782</TD><TD>8 7 8
                6 8 6</TD> </TR>
        </TABLEDATA> </DATA>
 </TABLE>
</RESOURCE>
</VOTABLE>
```

Рис. 1. Пример описания табличных данных с помощью VOTable 1.0

XML был выбран разработчиками по следующим причинам:

- включает в один документ данные и связанные с ними метаданные – описание содержимого;
- широко используется;
- может интерпретироваться, есть анализаторы инструментарий для работы;
- можно отображать информацию с помощью XSL;
- можно инкапсулировать в сообщения.

VOTable имеет синтаксис, соответствующий XML 1.0. Часть, описывающая данные, представлена спецификациями **<FIELD>** и **<PARAMETER>**. Описание метаданных должны включать следующую информацию:

Name	название колонки
Unit	единицы измерения
datatype	тип данных
Width	символьное представление
precision	точность
arraysize	размерность
Ucd	категория параметра

Данные, в виде XML-текста, FITS-файла, двоичного файла, вставляются между тегами **<DATA>**. Формат разрабатывался так, чтобы быть совместимым по типу данных, ключевым словам с FITS-форматом. Ключевые слова FITS-формата используются для семантического описания информации в файле. VOTable сохранил их использование в спецификации **<FIELD>**. Разработчики полагают, что VOTable может использоваться различным образом, как для передачи данных, так и для хранения, а также для хранения только метаданных. VOTable структура может пересылаться на сервер, который открывает высокоскоростное соединение, чтобы пересылать реальные данные, используя предварительно определенную структуру для интерпретации потока байтов через сокет. В качестве варианта, можно посылать только метаданные как неявный запрос серверу, который отвечает за наполнение структуры таблиц данными.

Для своей задачи, связанной с реализацией запросов одновременно к нескольким локальным архивам, были выбраны несколько типов наиболее часто используемых запросов: поиск по координатам, по имени объекта, по программе наблюдений.

Параметры файла одного из локальных архивов

Ключевое слово (FITS)	Значение	Комментарий
NAXIS1	530	Число пикселей 1 оси
NAXIS2	590	Число пикселей 2 оси
CDELT1	0.274	Приращение коорд. 1 оси
CDELT2	0.274	Приращение коорд. 2 оси
OBJECT	'TQ 66 '	Название объекта
OBSERVER	'Kaisin S.S. Tikhonov N.A.'	Наблюдатель
AUTHOR	'Karachentsev I.D. & Maslennokov'	Автор заявки
TELESCOP	'6m '	Телескоп
INSTRUME	'CCD1000 in the PF'	Прибор
DATE-OBS	'14/12/98'	Дата
TM-START	66785.	Начало экспозиции
EXPTIME	600.	Экспозиция
RA	24.4881666667	Прямое восхождение
DEC	15.4831944444	Склонение
PROG-ID	'Dwarf companions of spirals galaxies & Keuper belt'	Название наблюдательной программы
SEEING	2.	Качество изображения
FILTER	'1:R '	Фильтр

Для выполнения запросов в SIB для каждого файла вошли параметры, в качестве примера приведенные в таблице. Для каждого типа локального архива мы имеем почти совпадающие параметры, которые описывают целевую точку/область наблюдения, отличается описание этих параметров (ключевые слова, представление значений), а не физический смысл. С помощью VOTable описание локального архива, можно привести в соответ-

ствие ключевые слова разных локальных архивов, отвечающие, к примеру, за координатное представление, и выполнить запрос. При отображении результата запроса также используется VOTable-формат. В Страсбургском центре данных (CDS) имеется анализаторы и программное обеспечение для работы с этим форматом, которое мы предполагаем использовать в своей работе.

Литература

1. The National Virtual Observatory, Alexander S., Szalay, Astronomical Data Analysis Software and System X, ASP Conf. Ser., v.238, pp 3-12, 2001
2. The Astrophysical Virtual Observatory, http://www.eso.org/projects/avo
3. AstroGRID, http://www.astrogrid.ac.uk
4. The Russian Vurtuial Observatory. Principles and a Way of Realisation, V. Vitkovskij et al., Astronomical Data Analysis Software and System XI, in press
5. The Astrophysical Virtual Observatory. Drivers, Status and Planning, P. J. Quinn, in conf. “Toward an International Virtual Observatory”, Garching, 8-14 June, 2002 (in press).
6. Federation and Fusion of astronomical information. Standards and tools for the Virtual Observatories, D. Egret, F. Genova. in conf. “Toward an International Virtual Observatory”, Garching, 8-14 June, 2002 (in press). http://www.eso.org/gen-fac/meetings/vo2002/up/talks/egret
7. The GLU development site, http://simbad.u-strasbg.fr/glu/glu/htx
8. Definition of the Flexible Image Transport System (FITS), R.J. Hanish et al., A&A, v.376, pp 359-380, 1986
9. Применение FITS-формата для обмена и архивизации астрономических данных, В.В. Витковский, О.П. Желенкова, В.П. Рядченко, В.С. Шергин, Сообщения САО, т. 59, с. 60-67, 1988
10. The project of distributed information system OASIS, V. Vitkovskij et al., Baltic Astronomy, v.9, №4, pp 578-582, 2000
11. VOTable: A Proposed XML Format for Astronomical Tables, R. Williams et al., http://cdsweb.u-strasbg.fr/doc/VOTable