

Подход к разработке базы данных двойных звезд (BDB)*

Н. А. Скворцов¹, О.Ю. Малков², П. В. Кайгородов², Д. А. Ковалева²

¹Институт проблем информатики РАН, ²Институт Астрономии РАН
naskv@ipi.ac.ru, {malkov, pasha, dana}@inasan.ru

Аннотация

В статье представлен подход к разработке базы данных двойных звезд (Binary star DataBase, BDB). Она включает в себя информацию о физических и астрометрических параметрах двойных и кратных звезд различных типов, полученную из астрономических каталогов и обзоров. Рассматриваются требования к создаваемой базе, связанные с особенностями астрономических данных и объединением в одной базе данных информации из неоднородных информационных ресурсов. Описаны процессы формирования концептуальной схемы предметной области двойных звезд и преобразования данных каталогов к терминам концептуальной схемы. Предложены подходы решения научных задач в терминах концептуальной схемы предметной области.

1. Введение

За трехвековую историю астрономических наблюдений накоплено огромное количество наблюдательного материала, продолжают появляться новые данные. Обзоры переводятся в цифровую форму, развиваются средства наблюдения, появляются автоматизированные телескопы, в том числе, орбитальные. Обработанные данные наблюдений предстают в виде астрономических каталогов, представляющих собой упорядоченные коллекции большого количества данных определенного типа, собранных с определенной целью, и представляемых, как правило, в табличном виде.

Весьма востребованным видом наблюдательных данных являются данные о двойных звездах. Более половины, а по некоторым оценкам, более 90% звезд образуют двойные и кратные системы. Они могут быть обнаружены различными методами, и по этому признаку подразделяются на соответствующие классы: визуальные, фотометрические, спектроскопические, астрометрические, интерферометрические, рентгеновские и другие. Различные астрономические каталоги и базы данных обычно фокусируются на представлении части этих классов звезд. В частности:

- WDS (Washington Double Star catalogue) отражает данные о визуальных двойных;
- CEV (Catalog of Eclipsing Variables) – о затменных двойных;
- SB9 (9th Catalogue of Spectroscopic Binary Orbits) – о спектроскопических;
- INT4 (4th Catalog of Interferometric Measurements of Binary Stars) – об интерферометрических;
- CBS (Close Binary Systems) – о проэволюционировавших тесных двойных звездах.

Кроме того, имеется значительное число каталогов и баз данных, посвященных более узким выборкам двойных звезд, сформированных по различным, как физическим, так и наблюдательным критериям. Ни одна из существующих баз не включает в себя все типы двойных звезд. Все каталоги имеют собственную структуру таблиц, свой набор параметров о звездах, выделенных из разных обзоров неба разными методами.

Исследователям в области астрономии двойные звезды чрезвычайно важны для определения фундаментальных характеристик звезд: масс, радиусов и температур, – которые определяют их эволюцию. В частности, масса звезды не может быть напрямую определена из наблюдений одиночных звезд. Таким образом, двойные звезды – единственный источник данных для получения фундаментальных соотношений между звездными параметрами (соотношение "масса-светимость", соотношение "масса-радиус" и т.п.). Эти соотношения необходимы в астрономии для оценки параметров звезд, а также являются ключом к решению вопроса об образовании двойных и одиночных звезд. Помимо этого есть целый класс задач, решаемых с привлечением данных о двойных звездах, связанный как с изучением их самих, так и с более специальными интересами.

Таким образом, для решения научных задач астрономам приходится иметь дело с разрозненными каталогами, данные в которых могут содержать разные наблюдения и параметры одних и тех же астрономических объектов, а также оригинальных для конкретных каталогов объектов. Данные могут отличаться в результате различных условий и методов наблюдения, а также из-за различий в обработке и интерпретации результатов. Основной проблемой, с которой приходится сталкиваться при работе с астрономическими данными, является необходимость поиска объектов в большом количестве неоднород-

ных каталогов, объекты в которых плохо идентифицированы между собой.

База данных двойных звезд (BDB – Binary star DataBase) [3, 7] создается с целью заполнения существующего разрыва между каталогами двойных звезд и их пользователями, исследующими двойные звезды. Необходимо собрать каталогизированные данные о двойных звездах и предоставить возможность выборки звезд, отвечающих требованиям исследователя, и доступ к полному набору данных о них из разных каталогов, включенных в состав BDB.

Статья посвящена обоснованию решений при разработке базы данных двойных звезд. В следующем разделе статьи обсуждаются требования к создаваемой базе данных, исходящие из особенностей и нужд предметной области, связанной с их исследованием. В разделе 3 представлены обоснования подхода к проектированию базы данных. Четвертый раздел посвящен созданию концептуальной схемы предметной области двойных звезд, которая должна стать семантической основой для структурирования данных в создаваемой базе. Раздел 5 описывает принципы отображения схем астрономических каталогов в структуры концептуальной схемы предметной области. И в последнем разделе отражены функциональные возможности и особенности базы данных, созданной на основе построения спецификаций предметной области двойных звезд.

2. Особенности состояния предметной области и требования к создаваемой базе данных

Ключевой особенностью, отличающей астрономические данные от результатов физических экспериментов, является их уникальность, ценность каждого наблюдения. Это определяют несколько причин:

- измерения, проведенные в процессе наблюдений, не могут быть повторены с точным соблюдением исходных условий;
- наблюдения проводятся разными научными группами, разными методами, на различном оборудовании и с разными целями и набором исследуемых параметров;
- временные масштабы, на которых протекают явления, изучаемые астрофизикой, как правило, очень велики, поэтому астрофизические явления могут быть исследованы только на основе анализа наблюдений множества объектов, находящихся на разных стадиях одного и того же процесса;
- невозможность прямой экспериментальной проверки астрофизических гипотез требует тщательного сравнения их выводов с данными, накопленными за всю историю наблюдений.

Включаемые в BDB каталоги и базы данных весьма неоднородны. Можно выделить следующие причины их неоднородности:

- Большинство каталогов рассматривают в качестве основной единицы исследования пару звезд, входящих в кратную систему. Однако в некоторых единицах является компонент системы, то есть отдельная звезда, входящая в кратную систему. Причем в разных каталогах один и тот же объект может оказаться либо компонентом, либо парой, в зависимости от разрешения оборудования, используемого при наблюдении, и от специализации каталогов на определенных методах выделения кратных звезд.
- Нередки различия в идентификации компонентов систем, выделения главного компонента системы. В кратных системах со сложной структурой не все пары могут быть включены в исследование.
- Как уже упоминалось, различные цели составления каталогов приводят к тому, что они несут в себе различные наборы параметров, часто слабо пересекающиеся. Параметры могут быть представлены в неоднородном виде, имеют разные единицы измерения, измеряются разными методами с разной точностью.
- Помимо каталогов различных видов двойных звезд, в базу данных должны быть интегрированы каталоги астрофизических параметров, имеющих информацию о таких параметрах объектов, как масса, радиус, температура.

Особую проблему составляют идентификаторы звезд, пар и систем. С целью облегчения задачи поиска объекта по идентификатору значительная часть астрономических ресурсов предоставляет несколько (как правило, не менее двух) идентификаторов для каждого объекта. Однако значительное разнообразие типов астрономических идентификаторов делает проблему отождествления астрономических объектов во многих случаях весьма трудоемкой. Приходится прибегать к отождествлению объектов по координатам с привлечением дополнительной наблюдательной информации для исключения риска ложных отождествлений.

Еще одной проблемой является отслеживание изменений, поскольку с течением времени появляются новые наблюдения, которые могут как подтверждать результаты, полученные ранее, так и дополнять и уточнять их. При этом результаты любой обработки, проведенной для предыдущей версии данных, должны быть, соответственно, скорректированы.

Для решения многих астрономических задач требуется исследовать свойства определенных выборок астрономических объектов. Наборы требований к выборкам могут быть различными:

- объекты, определенным образом локализованные в пространстве (например, расположенные в ограниченном участке небесной сферы);
- объекты определенного диапазона блеска, с блеском в пределах чувствительности приборов или наблюдаемые с определенной точностью;

- объекты с одинаковыми или сходными наблюдательными характеристиками;
- объекты с определенной зависимостью между параметрами;
- объекты одного типа (например, переменные звезды, принадлежащие к определенному классу переменности)
- объекты, наблюдаемые в определенный период времени;
- пары с определенной пространственной ориентацией друг относительно друга или по отношению к наблюдателю;
- и другие.

Исходя из вышеупомянутых в данном разделе условий, связанных с предметной областью исследования двойных звезд и ее современным состоянием, можно составить основные требования, предъявляемые к создаваемой базе данных двойных звезд.

1. Принципиально важно учитывать уникальность каждого наблюдения, поэтому интеграция каталогов должна производиться без потери информации об отдельных измерениях, условиях их получения и каталоге, из которого взяты данные. Даже если несколько наблюдений из разных каталогов предположительно или точно связаны с одним астрономическим объектом, то каждое из этих наблюдений должно быть сохранено в базе данных.

2. Необходимо учитывать возможность пополнения базы новыми наблюдениями, интеграции новых каталогов, учета новых версий включенных каталогов. При этом результаты применения научных методов над базой должны быть скорректированы в учетом новых данных. Это значит, что разработанные и реализованные методы должны сохраняться вместе с базой и быть применимыми для обновляемых и новых данных.

3. Решение задач над базой данных является основным ее назначением. Необходимо обеспечить пользователям возможность производить выборку звезд по совершенно различным и, возможно, сложным критериям, вычисление параметров и соотношений по другим параметрам, предоставление в наиболее удобном для обработки виде всех данных, относящихся к определенным запрашиваемым астрономическим объектам.

4. В связи с принципами сохранения данных всех наблюдений и возможности пополнения базы новыми данными важно предоставить методы кросс-идентификации астрономических объектов (систем, пар, компонентов) по различным видам идентификаторов, встречаемых в каталогах, а также по координатам, с учетом наблюдательных параметров (блеска, собственных движений и других). Для существующих данных результаты работы этих методов должны быть сохранены в базе для обеспечения отождествления наблюдений одних и тех же объектов и предоставления исчерпывающих данных по определенным астрономическим объектам при решении задач.

3. Общий подход к разработке структуры базы данных двойных звезд

По определенным причинам, в числе которых основной является то, что в астрономии неповторимо и ценно каждое наблюдение, складываются характерные для этой предметной области подходы к решению задач с привлечением данных из различных астрономических каталогов.

Для решения типичных астрофизических задач необходимо проводить выборки звезд, обладающих определенным набором характеристик. Как правило, весь набор нужных характеристик не встречается в одном источнике и для решения задачи выбирается набор каталогов с необходимыми данными наблюдений. При этом выборе учитываются назначение каталогов и типы астрономических объектов, представленных в них, время их составления, точность измерений и условия наблюдения, состав наблюдаемых параметров, охват неба и другие возможные особенности. В выделенных каталогах по некоторым критериям производится выборка подходящих для решения задачи астрономических объектов. При этом искомым результатом будет пересечение выбранных из каждого каталога подмножеств (разных наблюдений одних и тех же объектов) необходима кросс-идентификация объектов между различными каталогами. Подобная кросс-идентификация обычно присутствует в составе самих каталогов, но зачастую, является неполной и для небольшой части объектов может быть ошибочной. Приходится также учитывать огромное разнообразие и неоднородность идентификаторов и методов идентификации. После того, как получена результирующая выборка, решаемые научные задачи формулируются в терминах конкретных выбранных каталогов. Задачи могут включать в себя сопоставление разных наблюдательных параметров из разных каталогов, вычисление через них других параметров или их отношений и построение диаграмм, которые позволяют установить искомые закономерности.

Такой подход, диктуемый особенностями предметной области и позволяющий сохранить в ней исходные данные наблюдений, тем не менее, обладает и своими недостатками. В первую очередь, он является слабо масштабируемым по количеству и составу используемых в решении задач каталогов. При появлении нового каталога или изменении структуры в новой версии каталога, программы, написанные для решения научных задач, придется переписывать. Каждая задача требует разработки решения в терминах нескольких конкретных каталогов. Всякий раз при решении задач приходится разбираться в структуре каждого каталога по отдельности.

Это влечет за собой и другой недостаток – невозможность повторного использования реализаций научных методов и решений задач. Программный код решения задач ориентирован на конкретные

каталоги, а не на методы, и каждый раз при решении задач над определенным составом каталогов астрономам приходится тратить немалые усилия на разработку реализации одних и тех же методов, включающих выборку данных из определенных каталогов, кросс-идентификацию объектов и собственно применение метода над данными каталогов.

Во избежание указанных недостатков в разработке базы данных двойных звезд предлагается идти от разработки спецификаций предметной области с учетом классов решаемых в ней задач. Такой подход включает в себя следующие этапы:

- разработку концептуальной схемы предметной области двойных звезд, предполагающей однородное представление если не всех, то преобладающего количества объектов и параметров;
- построение отображений схем конкретных информационных ресурсов (каталогов и баз данных) в схему предметной области с разрешением семантической неоднородности, связанной с разными единицами измерения или способами хранения данных;
- формулирование и решение задач в терминах концептуальной схемы предметной области, в том числе, построение узких надстроек над спецификациями предметной области для решения различных классов научных задач и персонализации ролей пользователей.

Такой подход позволяет избежать связи реализаций научных методов непосредственно с используемыми каталогами и разрабатывать их абстрактно в терминах спецификаций предметной области. Однако и в этом случае при разработке BDB необходимо учитывать перечисленные в предыдущем разделе требования, в особенности, обеспечение работы с уникальными наблюдениями без потери информации. Выработке соответствующих решений для обозначенных этапов разработки базы данных двойных звезд посвящены следующие разделы.

4. Разработка спецификаций предметной области исследования двойных звезд

Анализ предметной области начинается с взгляда на нее в целом:

- оценки ее охвата и границ, включаемых областей знания;
- описания возможного класса решаемых задач, методов их решения;
- составления текстов технического задания на разработку предметной области для решения некоторых классов задач;
- определения дополнительных источников знаний о предметной области;
- описания представительных информационных ресурсов (каталогов), характерных для предметной области.

Исходя из полученных описаний предметной области начинается разработка спецификаций. Подробнее подход описан в [6].

- Набирается глоссарий предметной области, состоящий из терминов и их определений, которые присутствуют в текстах.
- На основе терминов глоссария разрабатывается онтология предметной области. В ней определяются понятия, соответствующие терминам, связи и ограничения, соответствующие определениям терминов глоссария. На данном этапе учитываются знания общедоступных онтологий, разработанных в данной предметной области.
- На основе онтологии разрабатывается концептуальная схема предметной области, содержащая структуры для хранения и обработки данных, ограничения целостности, спецификации поведения: методы над объектами и процессы, происходящие в предметной области.

Для учета принципов сохранения уникальности наблюдений в спецификациях предметной области приняты следующие решения:

- Каждому идентифицируемому астрономическому объекту соответствует множество наблюдений. Именно отдельные наблюдения, а не идентифицируемые астрономические объекты рассматриваются в качестве объектов, соответствующих структурам в концептуальной схеме.
- Для сохранения информации о каталогах с каждым измерением необходимо хранить метаданные о каталогах, из которых они взяты. Метаданные могут содержать эпохи наблюдения, качество наблюдения, параметры оборудования, в частности, диапазоны фильтров, другие особенности каталогов и собственно названия каталогов. Описания каталогов становятся также объектами концептуальной схемы.
- Концептуальная схема предметной области должна включать спецификацию методов кросс-идентификации для связи объектов наблюдений, относящихся к одним и тем же астрономическим объектам.

При составлении онтологии предметной области исследования двойных звезд за основу структуры онтологии приняты знания, отраженные в онтологии астрономических объектов IVOA [1]. Однако этой онтологии астрономических объектов явно недостаточно для подробного описания предметной области двойных звезд. Были учтены описания требований к проекту создаваемой базы данных, описания некоторых каталогов двойных звезд и в них выявлены существенные понятия предметной области.

Рассмотрим фрагмент разработанной онтологии. Здесь и далее в примерах будет использоваться язык СИНТЕЗ, его подробное описание можно найти в [4]. Выбор этого языка для спецификаций обусловлен проектами по разработке спецификаций концептуальных схем предметных областей. Язык является языком спецификации, включающим фреймовую, объектную модели и язык формул, использующий многосортную логику предикатов первого порядка. Язык однородным образом используется

для спецификации онтологий и объектных концептуальных схем. Реализация же спецификаций может быть выполнена любыми средствами при наличии отображения языка СИНТЕЗ в соответствующие модели данных.

```
{ catalog; in: class;
  instance_section: {
    name: identifier;
    hasObject: {set_of: astrObject};
  }}
{ astrObject; in: class;
  instance_section: {
    inCatalog: catalog;
    metaslot inverse: catalog.hasObject end
    hasIdentifier: {set_of: identifier};
    hasCoordinate: coordinate;
    hasMeasurement: {set_of: measurement};
    hasProcess: {set_of: process};
    hasSpectralType: spectralType;
    hasLuminosityClass: luminosityClass;
    observedBy: discoverer;
    inEpoch: epoch;
    isComponentOf: {set_of: compoundObject};
    metaslot
      inverse: compoundObject.hasComponent
    end
  }}
{ compoundObject; in: class;
  superclass: astrObject;
  instance_section: {
    hasComponent: {set_of: astrObject};
  }}
{ stellarObject; in: class;
  superclass: astrObject;
  }}
{ multipleStar; in: class;
  superclass: compoundObject;
  instance_section: {
    multInv: { in: invariant;
      {{ all x, y ( multipleStar(x) & hasComponent(x, y)
        -> stellarObject(y)) }}
    }
  }}
{ multipleStarSystem; in: class;
  superclass: multipleStar;
  instance_section: {
    hasReferenceStar: component;
    hasPair: {set_of: pair};
  }
  }}
{ pair; in: class;
  superclass: compoundObject;
  instance_section: {
    ofSystem: system;
    metaslot inverse: system.hasPairs end
    hasPrimaryComponent: component;
    hasSecondaryComponent: component;
    hasObservationalType: {set_of:
      ObservationalType};
    hasEvolutionaryStatus: EvolutionaryStatus;
```

```
comp: {in: invariant; {{ all p (pair(p) ->
  p.hasComponent =
    {p.hasPrimaryComponent,
      p.hasSecondaryComponent} ) }} }
  }}
{ component; in: class;
  superclass: stellarObject;
  instance_section: {
    isReferenceStarOf: multipleStarSystem;
    metaslot
      inverse multipleStarSystem.hasReferenceStar
    end
    isPrimaryComponentOf: pair;
    metaslot inverse pair.hasPrimaryComponent end
    isSecondaryComponentOf: pair;
    metaslot inverse pair.hasPrimaryComponent end
  }}
}}
```

Фрагмент онтологии определяет следующие понятия:

- каталог (catalog).
- астрономический объект (astrObject);
- составной объект (compoundObject), состоящий из некоторого множества известных компонентов;
- звездный объект (stellarObject);
- кратная звезда (multipleStar) описывает визуально близкие звездные объекты;
- кратная звездная система (multipleStarSystem), являющиеся основным предметом исследования, включают только физически и эволюционно связанные звездные объекты;
- специфическое для изучения кратных систем понятие пары компонентов системы (pair) как подпонятие составного объекта;
- и компонент системы (component), являющийся звездным объектом.

В понятиях определяются их связи с другими понятиями и ограничения в виде инвариантов. Так, в описании понятия астрономического объекта определены связи с координатами, измерениями, физическими процессами и другими понятиями, также определенными в онтологии, но не приведенными в данном фрагменте. Помимо приведенных понятий в онтологии присутствуют понятия, относящиеся к описаниям координатных систем (coordinate, coordinateSystem, equatorialCoordinateSystem, rightAscension, declination), фотометрических систем (magnitude, passband, photometricSystem), к классификациям двойных звезд по наблюдательному и эволюционному критериям (ObservationalType, EvolutionaryStatus), к различным параметрам как общим для астрономических объектов, так и специфическим для звездных объектов, двойных систем и их разновидностей. Для спецификации набора наблюдательных и вычисляемых астрофизических параметров в соответствии с онтологией астрономических объектов IVOA введено понятие measurement, определенным параметрам соответствуют его подпонятия. Например, для астрономических объектов параметрами являются звездная величина (magnitude), соб-

ственное движение (*properMotion*), параллакс (*parallax*), возраст (*age*), масса (*mass*), температура (*temperature*), лучевая скорость (*radialVelocity*) и другие. Для описания двойных систем, а также пар внутри систем служат понятия период вращения (*period*), соотношение масс (*massRatio*), параметры положения (*positionAngle* и *angularSeparation*), параметры орбит (*semiMajorAxis*, *inclination*, *eccentricity*) для двойных систем. Отдельно описано понятие кросс-идентификации (*crossReference*) астрономических объектов.

При преобразовании онтологии в концептуальную схему предметной области части понятий ставятся в соответствие абстрактные типы данных, другие становятся типами перечисления, заменяются скалярными типами, преобразуются в функции, а некоторые – напрямую не используются и не входят в схему. Также сокращается иерархия понятий, и связи из неиспользуемых понятий переходят в подтипы. Связям понятий ставятся в соответствие атрибуты или ассоциации со значениями определенного типа.

При формировании структуры данных в астрономии простым решением было бы создание универсальной таблицы с множеством атрибутов, содержащих все возможные параметры. К этому подталкивает подобная табличная организация астрономических каталогов. Однако этот подход означает, во-первых, что при интеграции множества каталогов универсальная таблица будет содержать большое количество пустых полей в записях. Лучшим подходом является создание абстрактных типов данных для семантически связанных данных, либо в тех случаях, когда с одним параметром связан набор других.

Если в предметной области ожидается использование большого количества однотипных значений, связанных с одним физическим объектом, например, значений звездных величин в произвольных полосах пропускания, то в схеме предметной области целесообразно создать отдельный абстрактный тип данных для таких значений, в данном случае, тип звездных величин (*Magnitude*), содержащий связи с астрономическим объектом и полосой пропускания и значение звездной величины в данной полосе пропускания. Это позволяет сделать предметную область более структурированной и избежать раздутых объемов базы данных за счет хранения большого количества пустых полей.

С другой стороны, различные параметры, описывающие один и тот же физический объект, целесообразно хранить в значениях одного типа данных. В частности, понятия, являющиеся измерительными параметрами (подпонятия понятия *measurement*) преобразуются в атрибуты типов, соответствующих астрономическим объектам (парам, компонентам), чьи параметры они описывают. Параметры орбиты вращающихся двойных систем также преобразуются в отдельный абстрактный тип (*Orbit*).

В соответствии с описанной выше онтологией путем преобразования понятий и их связей в струк-

туры для представления данных и описания их поведения создается концептуальная схема предметной области. Приведем ее фрагмент.

```
{ Cat; in: class;
  instance_section: {
    name: String
  }
}
{ CatLine; in: class;
  instance_section: {
    catalog: Cat;
    number: integer;
  }
}
{ BDBSystem; in: class;
  instance_section: {
    cat: Cat;
    line: CatLine;
    id: String;
    pairs: {set_of: BDBPair};
    objects: {set_of: BDBComp};
    refObject: BDBComp;
    key: {unique; {cat, id}};
  }
}
{ BDBPair; in: class;
  instance_section: {
    cat: Cat;
    line: CatLine;
    system: BDBSystem;
    metaslot inverse: System.pairs end
    id: String;
    primary: BDBComp;
    secondary: BDBComp;
    crd: Coordinate;
    epoch: Integer;
    plx: Float;
    pm: ProperMotion;
    min: Magnitude;
    max: Magnitude;
    V0: Float;
    theta: Float;
    rho: Float;
    orbit: Orbit;
    oType: ObsType;
    eType: EvolStatus;
    key: {unique; {cat, system, id}};
  }
}
{ BDBComp; in: class;
  instance_section: {
    cat: Cat;
    line: CatLine;
    system: BDBSystem;
    metaslot inverse: System.objects end
    id: String;
    primaryOf: BDBPair;
    metaslot inverse: BDBPair.primary end
    secondaryOf: BDBPair;
    metaslot inverse: BDBPair.secondary end
    refObjectOf: System;
    metaslot inverse: System.refObject end
    crd: Coordinate;
    epoch: Integer;
```

```

    plx: Float;
    pm: ProperMotion;
    mag: {set_of: Magnitude};
    spType SpectralType;
    Rv: Float;
    K: Float;
    key: {unique; {cat, system, id}};
}
{ Coordinate; in: type;
  crdRA: Float;
  crdDE: Float;
}
{ ProperMotion; in: type;
  pmRA: Float;
  pmDE: Float;
}
{ Magnitude; in: type;
  value: Float;
  band: Passband;
}
{ Orbit; in: type;
  P: Float;
  A: Float;
  I: Float;
  E: Float;
}
{ crossRef; in: class;
  instance_section: {
    id: String;
    kind: enum;
    refs: {set_of: String}
    trans: {in: invariant;
      {{ all x/crossRef, y/crossRef, z/crossRef (
        in(x.refs, y.id) & in(y.refs, z.id) ->
        in(x.refs, z.id) ) }} }
  }
}

```

В приведенном фрагменте концептуальной схемы предметной области, составленной на основе преобразования из онтологии, определены основные классы для описания каталогов (Cat и CatLine), кратных звездных систем (BDBSystem), пар компонентов системы (BDBPair), компонентов (BDBComponent). Типы экземпляров этих классов определяют наборы их атрибутов (функции не приведены).

Из-за преобразования понятий в структуры данных часть знаний о предметной области при переходе от онтологии к схеме теряется. Чтобы избежать потери, используется семантическое аннотирование с помощью онтологии. Онтологию нужно сохранить в том виде, в каком она разработана, а каждый элемент концептуальной схемы связать с понятием онтологии, которому он соответствует, или с выражением, определяющим такое понятие в терминах онтологии.

В таблице 1 приведены примеры аннотации типа BDBPair и его атрибутов в терминах онтологии. Тип BDBPair соответствует понятию пары (pair). Его атрибуты cat, catLine, system, id, epoch и другие аннотируются соответствующими понятиями онтологии. А для атрибутов primary, secondary и crd со-

ставлены выражения, дополнительно ограничивающие понятия онтологии. В частности, аннотация атрибута crd означает, что в терминах онтологии он соответствует координатам объекта в экваториальной системе.

В дальнейшем эти аннотации понадобятся для идентификации онтологически релевантных элементов схем предметной области и конкретных информационных ресурсов.

Таблица 1. Аннотирование элементов схемы предметной области в терминах онтологии

Элемент	Понятие
BDBPair	pair
- cat	catalog
- catLine	record
- system	multipleStarSystem
- id	identifier
- primary	{x/component ex y (pair(y) & hasPrimaryComponent(x,y))}
- secondary	{x/component ex y (pair(y) & hasSecondaryComponent(x,y))}
- crd	{ x/coordinate all y (coordinateSystem(x,y) -> equatorialCoordinateSystem(y))}
- epoch	epoch
...	...

5. Отображение схем информационных ресурсов в концептуальную схему предметной области

Когда спецификации предметной области разработаны, необходимо произвести связывание концептуальной схемы с информационными ресурсами, данные из которых будут использоваться в разрабатываемой базе данных двойных звезд.

Построение отображения схем информационных ресурсов заключается в составлении правил отображения, выражающих классы концептуальной схемы через структурные элементы схем информационных ресурсов. Обычно такой подход плохо масштабируется по числу информационных ресурсов, так как правила отображения включают соединение (join) экземпляров из нескольких ресурсов и при появлении новых информационных ресурсов должны быть разработаны заново с учетом соединения с данными из них.

Однако особенность отображения в данном случае состоит в том, что каждое наблюдение, зафиксированное в каталоге, рассматривается в концептуальной схеме предметной области как отдельный объект. В результате, между записями соединения не производится. Идентификация объектов, описывающих одни и те же звезды, производится не на уровне правил отображения, а на уровне методов, определяемых в концептуальной схеме, и в общую базу для этого попадают все объекты из разных ресурсов. Следовательно, правила отображения над разными информационными ресурсами могут стро-

иться независимо друг от друга. (Исключение составляют ссылки каталогов друг на друга, например, каталог CCDM использует (поле Disc) коды первооткрывателей, включённые в каталог WDS, а также ссылки (поле Note1) на параметры орбит во внешнем источнике. Эти источники желательно интегрировать одновременно с каталогом.) Подход к отображению остается масштабируемым по количеству и изменению состава ресурсов.

Если для интеграции новых информационных ресурсов (новых каталогов или новых версий интегрированных каталогов) необходимо развитие самой концептуальной схемы предметной области, то расширяется онтология предметной области новыми понятиями и связями без изменения прежних, и обновление схемы должно выполняться как дополнение новыми классами с типами экземпляров и новыми атрибутами существующих типов. Отображение структур новых ресурсов в концептуальную схему предметной области производится с учетом уже существовавших в ней структур и новых элементов. При таком подходе, данные, уже присутствующие в базе данных, дополняются пустыми полями, и других изменений с ними не происходит, а данные из новых ресурсов формируют новые экземпляры в базе данных. Подход является масштабируемым для обновлений схемы в соответствии с новыми ресурсами.

Интеграция информационных ресурсов в базу данных двойных звезд может быть выполнена двумя способами: виртуальным или материализованным. При виртуальной интеграции данные не переносятся из ресурсов. Запросы к концептуальной схеме предметной области переписываются в термины схем информационных ресурсов с помощью GAV-взглядов [5], составленных в соответствии с правилами отображения. Ответы на запросы к каталогам преобразуются в термины предметной области и возвращаются в качестве ответа на запрос к концептуальной схеме предметной области.

Материализованная интеграция предполагает проведение ETL-процесса (Extraction-Transformation-Load). Необходимо извлечь (extract) каталоги, представив их локально в исходных файлах или в реляционной базе данных в их исходной структуре, преобразовать (transform) данные в соответствии с правилами отображения к структурам, соответствующим концептуальной схеме предметной области, и, сформировав запрос к концептуальной схеме, загрузить (load) их в базу данных, построенную в соответствии со структурой предметной области.

Более подробно рассмотрим подход к отображению схем и преобразованию данных на примере отображения схемы каталога CCDM [2] в концептуальную схему предметной области. Каталог состоит из одной основной таблицы и одной вспомогательной, их структура должна быть представлена в виде схемы ресурса. Фрагмент схемы ресурса CCDM выражен на языке СИНТЕЗ для преодоления модельной неоднородности спецификаций ресурса и предметной области:

```
{ CCDMRec; in: class
instance_type: {
  recno: Integer;
  CCDM: String;
  rComp: Byte;
  Comp: Byte;
  dRAs: Float;
  dDEs: Float;
  Year: Integer;
  Theta: Float;
  Rho: Float;
...
}}
```

Приведены лишь некоторые поля основной таблицы: CCDM – идентификатор системы звезд, составленный на основе ее базовых координат; comp – алфавитный код текущего компонента в системе; rComp – код компонента, с которым данный состоит в паре; dRAs и dDEs – относительные координаты компонента в системе; Year – год его светимости в видимом диапазоне. Theta и Rho – положение звёзд

Связывание схемы ресурса со спецификациями предметной области начинается с семантического аннотирования элементов схемы понятиями онтологии предметной области. Коды UCD, известные для полей каталога, могут оказаться полезными для такого связывания.

Таблица 2. Аннотирование элементов схемы каталога в терминах онтологии

Элемент	Понятие
CCDMRec	component
Recno	Record
- CCDM (meta.id; meta.main)	multipleStarSystem, identifier, { x/coordinate all y (coordinate- System(x,y) -> equatorialCoordinateSystem(y)) }
- rComp (meta.code. multip)	component identifier
- Comp (meta.code. multip)	component identifier
- dRAs (pos.eq.ra; arith.diff)	rightAscension
- dDEs (pos.eq.dec; arith.diff)	declination
- Year (time.epoch)	Epoch
- Theta (pos. posAng)	positionAngle
- Rho (pos. angDistance)	angularSeparation
...	...

В приведенных в таблице 2 аннотациях класс CCDMRec соответствует понятию компонент (component), атрибут CCDM описан в терминах онтологии с одной стороны как ссылка на систему, с другой как идентификатор компонента, и к тому же, как координата объекта. rComp и Comp в совокупности с полем CCDM составляет CCDM-идентификатор компонента системы. Атрибуты dRAs и dDEs аннотируются понятиями координат, так как в онтологии нет лучших средств описания понятия смещения координат. Год измерения позиции Year соответствует эпохе и так далее.

Следующим этапом является идентификация онтологически релевантных элементов схемы ресурса и схемы предметной области. Он производится на основе решения задачи включения между классами семантических аннотаций схем и выделения соответствий между понятиями, определяемыми аннотациями. Онтологически релевантные элементы схем выбираются среди аннотаций, имеющих соответствия и подтверждаются либо отклоняются экспертом. В таблице 3 приведены релевантные элементы схем:

Таблица 3. Онтологически релевантные элементы схем предметной области и ресурса

Элемент ресурса	Элемент схемы предметной области
CCDMRec	BDBComp
- Recno	CatLine.number
- CCDM	BDBSystem.id BDBComp.crd
- rComp	BDBPair.id BDBPair.primary
- Comp	BDBPair.id BDBPair.secondary BDBComp.id
- dRAs	BDBComp.pmRA
- dDEs	BDBComp.pmDE
- Year	BDBPair.epoch
- Theta (pos. posAng)	BDBPair.theta
- Rho (pos. angDistance)	BDBPair.rho
...	...

Онтологическая релевантность элементов схем используется для определения того, значения каких элементов схемы ресурса должны использоваться для формирования значений определенного элемента концептуальной схемы предметной области при отображении. Некоторые релевантные элементы схем отображаются друг в друга в виде присвоения значения одного элемента другому. Например:

BDBSystem.id = CCDMRec.CCDM
BDBPair.epoch = CCDMRec.Year

Если между элементами схем присутствует структурная неоднородность какого-либо вида (необходимость вычисления через другие элементы, преобразования единиц измерения, прохода по пути атрибутов и другие), спецификация отображения элементов схемы ресурса в элементы концептуальной схемы предметной области выражается в виде функций разрешения структурных конфликтов.

```
{ ccdmGetCompId; in: function;
  params:
    (– ccdm/String, – comp/String, +return/String);
  {{ return = ccdm + comp }}
}
```

Например, функция разрешения конфликта для элемента BDBComp.id формирует идентификатор из двух полей схемы ресурса. Подобные функции должны быть построены для всех элементов, не имеющих примитивного отображения.

И наконец, построение правил отображения производится на основе соответствий элементов схем с использованием функций разрешения конфликтов между ними.

```
Cat(x/[name]):=
  name='CCDM'
```

```
CatLine(x/[number]):=
  CCDMRec(y/[Recno]) &
  number = RecNo
```

```
BDBSystem(x/[cat, line, id]) :-
  CCDMRec(y/[Recno, CCDM, rComp, Comp,
    dRAs, dDEs, Year, Theta, Rho, pmRA, pmDE]) &
  Cat(c/[name:cat]) &
  CatLine(l/[number:line]) &
  cat = 'CCDM' &
  line = RecNo &
  id = CCDM &
  Comp = 'A'
```

```
BDBPair(x/[cat, line, system, id, primary, secondary,
epoch, theta, rho]) :-
  CCDMRec(y/[ Recno, CCDM, rComp, Comp,
    dRAs, dDEs, Year, Theta, Rho, pmRA, pmDE]) &
  BDBSystem(system/[cat, id:CCDM]) &
  BDBComp(primary/[cat, system, id:primId]) &
  BDBComp(secondary/[cat, system, id:secId]) &
  Cat(c/[name:cat]) &
  CatLine(l/[number:line]) &
  ccdmGetPairPrimaryId(CCDM,rComp,primId]) &
  ccdmGetPairSecondaryId(CCDM,Comp,secId]) &
  ccdmGetPairId(CCDM, rComp, Comp, id) &
  epoch = Year &
  theta = Theta &
  rho = Rho &
  cat = 'CCDM' &
  line = RecNo &
  rComp != ''
```

```

BDBComp(x/[cat, line, system, id, refObjectOfSystem,
  crd, pm]) :-
  CCDMRec(y/[ Recno, CCDM, rComp, Comp,
    dRAs, dDEs, Year, Theta, Rho, pmRA, pmDE]) &
  Cat(c/[name:cat]) &
  CatLine(l/[number:line]) &
  ccdmGetCompId(CCDM, Comp, id) &
  ccdmGetCompRefObjectOfSystem
    (CCDM, refObjectOfSystem) &
  ccdmGetCompPm(pmRA, pmDE, pm) &
  ccdmGetCompCrd(CCDM, dRAs, dDEs, pm) &
  cat = 'CCDM' &
  line = RecNo

```

Приведенные правила представляют собой фрагменты правил формирования экземпляров классов BDBComp, BDBPair, BDBSystem, CatLine и Cat. Каждое из правил использует класс ресурса (CCDMRec), функции разрешения конфликтов, присвоения и сравнения значений атрибутов. В правиле BDBPair, задающем формирование объектов пар, используются обращения к правилам отображения BDBComp для связи с объектами компонентов, входящих в пару, и BDBSystem для ссылки на систему. Также BDBComp использует правило отображения BDBSystem для задания атрибута главной звезды в системе.

Таким образом, построено отображение (фрагмент отображения) схемы каталога CCDM в спецификации предметной области создаваемой базы данных двойных звезд. Далее можно приступить к формированию базы данных путем ETL-процесса. Вначале производится пробное распознавание и преобразование на запросе, ограниченном первыми ста ответами от каталога. Полное извлечение, преобразование данных в термины предметной области базы данных BDB и загрузка их в базу происходят автоматически с помощью запроса к концептуальной схеме. Ошибки распознавания должны быть проанализированы экспертом. Процедура интеграции любых других каталогов будет подобна интеграции CCDM.

6. Интерфейсы базы и решение научных задач над ней

Текущая реализация базы данных BDB содержит в описании схемы предметной области те параметры, по которым предполагается организовать поиск данных. Отображены и загружены ряд каталогов, включая основные каталоги для визуальных двойных звезд. Интерфейс системы содержит поиск по идентификаторам разных форматов, выдачу информации о составе систем звезд в разных каталогах с учетом кросс-идентификации. С использованием Google реализовано изображение систем на звездном небе. На данный момент реализована возможность просмотра объектов в оригинальной структуре таблиц каталогов.

Описанный в статье подход к формированию предметной области базы данных и интеграции в нее информационных ресурсов позволит построить общую схему для изучения двойных звезд и отображать в нее возможно большее подмножество параметров интегрируемых каталогов. Таким образом, решение задач можно было бы осуществлять полностью в терминах предметной области базы данных, не прибегая к схемам исходных каталогов. В этом случае в интерфейсе базы данных можно будет:

- осуществлять поиск объектов по сложным критериям;
- задавать запросы к структуре предметной области, ограниченные только возможностями реляционного языка запросов;
- писать программы над структурой предметной области.

Независимость от состава интегрированных ресурсов заставляет формулировать задачи не в требованиях к каталогам, а абстрактно, в терминах предметной области. В этом случае рациональным становится:

- накопление знаний о предметной области, выраженных в терминах онтологии и концептуальной схемы;
- создание коллекции реализованных научных методов;
- создание специализированных надстроек (data mart) над спецификациями предметной области, содержащих ограниченные интерфейсы для решения определенных классов задач и рассчитанных на специалистов узкого профиля.

Одна из задач является неотъемлемой частью создаваемой базы данных. После загрузки данных в базу присутствуют столько записей о каждом астрономическом объекте, сколько информационных ресурсов содержали информацию об этом объекте. Задачей экспертов, работающих в терминах концептуальной схемы базы данных является кросс-идентификация записей об одних и тех же объектах. Спецификации концептуальной схемы BDB содержат спецификации процесса кросс-идентификации данных, загруженных из разных информационных ресурсов. Он включает учет соответствий множественных указаний идентификаторов, которые присутствовали в самих ресурсах. Эти соответствия также имеют свойство транзитивности, что учтено в схеме. Соответствия могут содержать и ошибки, для обнаружения и исправления которых в схеме должны быть также разработаны алгоритмы. Другой подход к кросс-идентификации, который должен применяться в базе данных – отождествление по близости координат, собственных движений, светимостей и других параметров. Задача кросс-идентификации и результаты ее работы будут использоваться на всех этапах функционирования BDB. Поэтому целесообразна реализация кросс-идентификации и как набора знаний о предметной области, и как общедоступной реализации набора научных методов в виде сервисов над базой данных, и как

специализированный интерфейс эксперта над концептуальной схемой базы данных BDB.

7. Заключение и дискуссия

В статье предложен подход к разработке базы данных двойных звезд, при котором создается спецификация предметной области, в нее отображаются интегрируемые астрономические каталоги, и решение научных задач предполагает их описание в терминах спецификаций предметной области. Таким образом, научные методы формулируются единожды исчерпывающим образом в терминах предметной области. Возможно повторное использование реализованных научных методов и решения задач.

Обеспечивается масштабируемость системы по количеству и изменению состава участвующих каталогов. Новый каталог отображается в схему предметной области и может участвовать в решении задач, сформулированных в терминах предметной области. Обеспечивается возможность обновления данных. При обновлении производится пересчет задач с новыми данными с использованием уже накопленных коллекций программ.

Вопросом для обсуждения остается возможность и целесообразность полного отображения всех параметров интегрируемых каталогов в структуру предметной области базы данных BDB. На момент написания статьи структура BDB включает основные параметры, по которым обычно осуществляется поиск двойных звезд, при этом дополнительные данные о звездах доступны только из оригинальных таблиц каталогов. Представляется целесообразным включать в схему BDB большее количество параметров. Это усложняет процесс разработки базы и интеграции каталогов, однако дает указанные выше преимущества при решении научных задач над базой. Необходимо найти разумный консенсус в представлении тех данных, которые когда либо могли бы быть использованы при решении задач не только для поиска, но и для вычислений и оценки параметров.

Литература

- [1] Ontology of Astronomical Object Types. Version 1.3. IVOA Technical Note, 17 January 2010. – IVOA. – 2010.
- [2] Dommanget J., Nys O. Catalogue des Composantes d'Étoiles Doubles et Multiples, Deuxième Édition. // Observations et Travaux. – 54, 5. – 2002.
- [3] Kaygorodov P., Debray B., Kolesnikov N., Kovalyova D., Malkov O. The new version of Binary star database (BDB) // Baltic Astronomy. – 21, N3. – 2012. – in press
- [4] Kalinichenko L.A., Stupnikov S. A., Martynov D.O. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Infor-

mation Resource Environments. – Moscow: IPI RAN, 2007. – 171 p.

- [5] Ullman J. D. Information Integration Using Logical Views // Proceedings of the 6th International Conference on Database Theory. – P. 19-40. – 1997.
- [6] Вовченко А.Е., Захаров В.Н., Калинин Л.А., Д. Ю., Рябухин О. В., Скворцов Н. А., Ступников С. А. От спецификаций требований к концептуальной схеме. // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», RCDL'2010. – Казань: КГУ. – 2010.
- [7] Малков О., Кайгородов П., Облак Э., Дебрэ Б. БДБ: база данных о двойных звездах. // Динамика сложных систем, 4, №2, 48. – 2010.

An approach to development of the Binary Star DataBase (BDB)

N. A. Skvortsov, O. Yu. Malkov,
P. V. Kaygorodov, D. A. Kovalyova

The paper presents an approach to development of the Binary star DataBase (BDB). The database contains data about physical and astrometric parameters of double star of different types, taken from astronomical catalogs and surveys. Requirements to the database are considered, which are coming from peculiarities of astronomical data and from the task of merging of information from heterogeneous information resources in a single database. The development process of conceptual scheme of binary stars subject area and transformation of catalog data to the conceptual scheme are described. An approach to scientific problem solving in terms of a conceptual scheme is proposed.

* Работа выполнена при поддержке грантов РФФИ 10-02-00426, 10-07-00342, 11-02-00076, 11-07-00402, 12-02-00047, 12-07-00528, при поддержке Федерального агентства по науке и инновациям (грант 02.740.11.0247), Программы Президиума РАН (программа 16П, проект 4.2), «Поддержка ведущих научных школ» (грант НШ-3602.2012.2), а также в рамках реализации ФЦП "Научные и научно-педагогические кадры инновационной России" на 2009 - 2013 годы.