Интеллектуальные системы управления, анализ данных

© 2025 г. Н.А. СКВОРЦОВ (nskv@mail.ru)
(Федеральный исследовательский центр
«Информатика и управление» РАН, Москва)

УПРАВЛЕНИЕ КАЧЕСТВОМ ДАННЫХ ПРИ РЕШЕНИИ ЗАДАЧ В ИССЛЕДОВАТЕЛЬСКИХ ИНФРАСТРУКТУРАХ НАД НЕОДНОРОДНЫМИ ИСТОЧНИКАМИ ДАННЫХ

Решение задач на основе доступных научных данных, особенно в контексте открытой науки и исследовательских инфраструктур, должно обеспечивать возможность их многократного повторного использования. Показатели качества данных являются важными характеристиками, влияющими не только на точность методов при решении исследовательских задач, но и на оценку пригодности данных, возможность решения конкретных научных задач, выбор методов работы с данными, их совместимость, возможность отождествления объектов и другие аспекты повторного использования. При этом требуется оценка различных показателей качества данных на разных уровнях агрегации – от целых наборов данных до отдельных значений. В данном исследовании представлен подход к комплексному управлению качеством данных на основе их спецификаций, а также требований к качеству данных и метаданных. Обсуждаются различные показатели оценки качества данных, включая точность, полноту и происхождение. Разработанный подход применен на примере решения задач с использованием множественных источников данных в области звездной астрономии.

Ключевые слова: качество данных, повторное использование данных, формальные спецификации, нефункциональные требования.

DOI: 10.31857/S0005231025040057, **EDN:** CARLWN

1. Введение

Научные исследования неизбежно сталкиваются с необходимостью оценки качества доступных данных, которые используются в процессе решения задач. На разных этапах научного исследования возникает потребность в оценке качества данных. При поиске и выборе подходящих наборов данных важно оценить их применимость, основываясь на информации о качестве данных в этих наборах. При создании выборок, включающих только подходящие данные, а также при очистке данных и их улучшении с целью подготовки к исследованию необходимо оценивать качество данных, связанных с конкретными объектами или характеристиками. При использовании научных методов важно учитывать качество исходных данных как для оценки качества

получаемых результатов, так и для оценки качества самих применяемых методов. Таким образом, показатели качества данных являются важным и даже необходимым элементом при решении различных исследовательских задач. Отсутствие информации о качестве данных или ее игнорирование может существенно снизить качество результатов, вплоть до их ошибочности, или сделать проведение исследования невозможным.

Во многих дисциплинах объем научных данных неуклонно растет. В условиях неоднородности источников научных данных и их разнообразия возникают проблемы совместимости данных, полученных разными методами, созданных с различными целями и имеющих разные требования к качеству.

В одних источниках информация о качестве представленных данных может содержаться в описании наборов данных, сопроводительной документации или быть известной из внешних исследований. В других источниках данные сопровождаются оценками качества, включенными в структуру самих наборов. В некоторых случаях информация о качестве данных отсутствует, но его можно оценить с помощью статистических методов, анализируя сами данные или их выборки. Методы оценки качества данных также могут существенно различаться. В зависимости от решаемых задач важными могут быть разные виды показателей качества и различные критерии их оценки. В некоторых случаях необходимо по-разному оценивать качество решений, исходя из информации о качестве исходных данных. Таким образом, неоднородность наблюдается и в самих метаданных, касающихся качества данных, и в их применении.

Для управления научными данными и поддержки исследований создаются исследовательские инфраструктуры, которые аккумулируют данные, предоставляют сервисы и метаданные, обеспечивая их повторное использование. Ввиду своей важности подходы к управлению качеством данных в исследовательских инфраструктурах требуют тщательных исследований. При работе с большими объемами данных и множеством неоднородных источников необходимо стремиться к автоматизированному управлению их качеством. Поэтому метаданные, касающиеся качества данных, должны быть четко определены, доступны и понятны как для человека, так и для машины.

В данной статье ставится задача разработки представления и применения различных показателей качества данных в исследовательских инфраструктурах. В следующем разделе представлен анализ состояния исследований в данной области. В разделе 3 предложена классификация подходов к управлению качеством данных. Описаны спецификации метаданных и принципы их использования в исследовательских инфраструктурах. Раздел 4 содержит пример управления качеством данных при решении задач в звездной астрономии с использованием данных больших многоцветных фотометрических обзоров неба. Сделаны общие выводы о развитии исследовательских инфраструктур для повышения эффективности научных исследований.

2. Состояние исследований в области управления качеством данных

Проблемы качества столь чувствительны в национальной и глобальной экономике, что для их решения давно прорабатывались подходы, методы и стандарты. В информатике вопросы качества данных, как часть общей проблемы качества объектов реального мира, играют столь же немаловажную роль. Базы данных и информационные системы рассматриваются как самостоятельные объекты [1], которые, как и другие объекты реального мира, могут иметь допустимое или недопустимое состояние, которое может оцениваться показателями качества данных. При этом данные в них отражают состояние объектов реального мира, и необходимо оценивать как соответствие состояния информационных объектов характеристикам объектов реального мира, так и возможность идентификации объектов реального мира по данным. Игнорирование проблем качества данных может привести к серьезным последствиям не только в информационной сфере, но и негативно сказаться на других сферах человеческой деятельности.

Оценка качества объектов реального мира основывается на данных об их качественных и количественных характеристиках. Таким образом, качество объектов (изделий, материалов и др.) всегда связано с качеством данных о них. Более того, в современной научной парадигме, основанной на извлечении новых знаний из научных данных, сбор большого объема данных разного качества об исследуемых объектах из различных источников требует оценки и учета их качества в исследованиях, что, в конечном счете, определяет развитие науки.

К данным, используемым в человеческой деятельности, могут предъявляться требования. Качество данных (такие характеристики, как точность и достоверность) об объектах реального мира может определять применимость и особенности использования как самих данных, так и объектов исследования. Низкое качество данных, даже если сами исследуемые объекты обладают высоким качеством, может затруднить решение задач и помешать правильному использованию как данных, так и объектов исследования.

С давних пор преобладающим подходом в оценке качества данных стали многокритериальные модели, которые включают наборы показателей качества. Уже в ранних исследованиях, таких как [2], обсуждается важность и востребованность различных показателей качества данных. Среди основных выделяются такие показатели, как точность, полнота, целостность и актуальность, а также достоверность, недвусмысленность, надежность и объем данных. При этом набор используемых показателей зависит от поставленных целей в конкретной предметной области исследования. В некоторых работах подчеркивается связь качества данных с качеством продукции и объектов реального мира, что подтверждает эффективность таких подходов к оценке качества данных. С этой точки зрения ценность данных определяется их потребителем, что обращает внимание на такие показатели качества, как значимость и корректность данных [3].

В процессе исследований устоялся подход, в котором модель качества определяется как пространство показателей качества, рассматриваемых в виде измерений. В [4] методология оценки качества определяется моделью, включающей определенный набор показателей, методы их оценки и, возможно, методы их обобщения. В каждой конкретной модели в зависимости от решаемых задач могут использоваться различные методы оценки показателей качества, такие как алгоритмы, правила, эвристики или модели машинного обучения, обеспечивающие решение задач, относящихся к определенным аспектам качества данных.

Так, например, полнота данных может быть оценена как полнота кортежей (заполнение всех атрибутов), атрибутов (оценка количества отсутствующих значений среди значений атрибута в кортежах) или как покрытие всех существующих объектов данного типа. Измерение точности может определяться как погрешность значений, синтаксическая точность (совпадение используемых имен для одного объекта) или семантическая точность (непротиворечивость фактов). Объем данных может определяться занимаемым объемом памяти или ожидаемым количеством возвращаемых кортежей и т.д. Другими словами, методы и реализации оценки качества могут зависеть как от характеристик наборов данных или выборок, так и от определенных ограничений предметной области, к которой относятся данные. Семантика того или иного показателя, несмотря на одинаковые названия, может существенно различаться в разных задачах. Соответственно требования к качеству данных могут формулироваться в исследовательских задачах как в терминах наборов данных или хранилищ данных, так и в терминах предметной области прикладной задачи.

Похожие подходы определены в действующих стандартах качества данных. С управлением качеством данных связаны международные и соответствующие им российские национальные стандарты, в частности ГОСТ Р 56214-2014 [5] (ISO/TS 8000-1: 2011) и комплекс сопутствующих стандартов. Эти стандарты соответствуют устоявшимся принципам моделирования качества данных и вводят такие показатели качества, как точность, полнота (или завершенность) и происхождение (или источник) данных, при этом не ограничивая конкретные реализации методов оценки качества по этим критериям. Кроме того, стандарты определяют концептуальные схемы управления качеством и принципы разработки критериев для оценки этих показателей, включая принципы однозначного синтаксического и семантического кодирования. Качество данных в рамках данных стандартов указывает на степень, в которой данные удовлетворяют требованиям потребителей и соответствуют установленным критериям. Устанавливаются принципы предъявления требований и повышения качества данных, позволяющие потребителям запрашивать данные надлежащего качества и точно определять, соответствуют ли полученные данные установленным стандартам.

Интересен также стандарт ГОСТ Р 57773-2017 [6] (ISO 19157:2013), который закрепляет практику управления качеством при работе с источниками пространственных данных, имеющими разное разрешение и детализацию. Этот стандарт позволяет выбирать наиболее подходящие по качеству данные для решения задач в требуемых масштабах.

Проблемы доступности научных данных, подходящих для исследований, возникают во многих областях с интенсивным использованием данных. В связи с этим активно развивается направление исследовательских инфраструктур, которые объединяют исследовательские данные, сервисы и инструменты, позволяя многократно их использовать.

Одной из идеологических основ их развития стали руководящие принципы FAIR-данных [7], которые предлагают направление для обеспечения обнаруживаемости (Findability, F), доступности (Accessibility, A), интероперабельности (Interoperability, I) и, как следствие, возможности повторного использования (Reusability, R) исследовательских данных. При разработке исследовательских инфраструктур и принципов их работы вопросы обеспечения качества данных часто сводятся к подходам оценки хранилищ данных с точки зрения соответствия принципам FAIR.

Крупные междисциплинарные инициативы научного сообщества, направленные на обеспечение и управление данными, а также на их анализ, задаются вопросами обеспечения качества данных, в основном, с позиции принципов FAIR, смещая акцент с управления качеством данных на оценку качества инфраструктуры управления данными для возможности улучшения качества данных.

В инициативе RDA разработана модель FAIR Data Maturity Model [8], которая определяет набор показателей, их приоритеты и методы для оценки следования принципам FAIR. Она используется в качестве общего подхода для оценки различных методологий. В частности, проект FAIRsFAIR [9] был создан для популяризации принципов FAIR для данных, создаваемых исследователями. Были разработаны популярные разъяснения, примеры и варианты решений, которые способствуют интероперабельности и повторному использованию данных, а также инструменты, стандарты и практики для обеспечения управления данными в различных научных дисциплинах. На базе модели FAIR Data Maturity Model и разработок FAIRsFAIR реализован инструмент F-UJI [10] для оценки степени соответствия принципам FAIR научных наборов данных и предоставления рекомендаций. Таким образом, подобные разработки создают альтернативный подход к оценке качества данных. Оценка следования принципам FAIR дает осязаемую оценку качества управления данными, а не самих данных, в основном. С другой стороны, это может обесценить сами принципы FAIR как некоторое стратегическое направление в развитии управления данными. Ведь эти принципы остаются перспективными, пока в достаточной мере невыполнима декларируемая ими возможность автономного управления дан-

ными машиной (machine-actionability), т.е. переход от обработки машиной предопределенных наборов метаданных и инструкций к обеспечению корректной интерпретации данных и метаданных, с которыми машина ранее не работала.

Инициатива ELIXIR специализируется на биомедицинских и биологических информационных ресурсах. Здесь вопросы качества данных также чаще обсуждаются именно в контексте соответствия принципам FAIR, в частности в рекомендациях FAIR CookBook [11]. Однако специальных рекомендаций об управлении качеством данных не выработано.

Исследовательская инфраструктура ELIXIR [12, 13] (в частности, платформы Data и Interoperability в ее составе) поддерживает хранение больших объемов данных, развитое управление метаданными, контроль целостности данных, механизмы обнаружения дубликатов и данных и другие средства, используемые для управления качеством данных. При этом специализированных средств для спецификации качества данных, помимо оценки выполнения принципов FAIR, не предусмотрено.

В рамках инициативы ESIP, посвященной управлению данными в области наук о Земле, проводились исследования, связанные с управлением качеством данных. Рабочие группы Data Quality Working Group (DQWG) и Information Quality Cluster (IQC) разработали рекомендации применения лучших практик и стандартов в этой области, принципов управления жизненным циклом данных для обеспечения их целостности, качества и повторного использования, а также способствовали внедрению этих рекомендаций среди провайдеров данных, разработчиков программных систем и научных групп [14–16]. Эти решения во многом связаны со сложившейся практикой работы с пространственными данными, такими как стандартизация атрибутов и метаданных, использование флагов качества в данных, следование специализированным стандартам.

Также в рамках ESIP разработана матрица оценки зрелости научных данных, объединяющая в себе основные принципы курирования данных при долгосрочном хранении и эксплуатации. Частью этих принципов является управление качеством данных [17].

Международные стандарты управления качеством данных также продолжают развиваться. Комплекс стандартов управления качеством пространственных данных находится в процессе обновления [18].

Большинство распространенных продуктов, связанных с управлением качеством данных, является инструментами очистки данных в составе систем интеграции. Задачами, решаемыми в таких продуктах, являются выявление ошибок и дубликатов данных, приведение типов и стандартизация представления значений, решение проблемы отсутствующих данных и др. Однако такие продукты также могут предоставлять средства оценки качества данных и варианты алгоритмов вычисления показателей качества. К таким продуктам относится, в частности, Talend Data Quality [19], являющийся частью систе-

мы интеграции реляционных баз данных. В этом инструменте используются бизнес-правила на основе SQL-запросов для представления требований к данным и мониторинга качества данных по таким показателям, как полнота, точность и согласованность данных с возможностью выявления конкретных проблемных областей. IBM InfoSphere Quality Stage [20], входящий в линейку продуктов IBM для поддержки процессов управления данными, обеспечивает непрерывный мониторинг событий качества в потоках данных, исправление ошибок, очистку данных, обогащение метаданных.

В основе Semantic Data Quality Management (SDQM) [21] лежит анализ стандартов ISO 9000 и исследования моделей качества данных с наборами показателей, а также технологии семантического Веба [22], включая RDF-технологии [23]. Семантические подходы на основе RDF, языка с расширяемой семантикой для описания ресурсов, позволяют связывать с ресурсами, идентифицируемыми в глобальном информационном пространстве, метаданные, семантика которых определяется словарями или онтологиями в различных пространствах имен, задавать требования к ресурсам в терминах этих словарей и находить релевантные ресурсы. Более того, модель RDF стала наиболее часто используемой для определения схем метаданных. Для определения модели качества вводится словарь управления качеством данных [24], повторно используются модель происхождения данных для веб-ресурсов [25] и другие словари.

Big Data Quality Management Framework (BDQMF) [26] предназначен для управления качеством больших данных на всех этапах их жизненного цикла. BDQMF охватывает весь процесс управления качеством данных: от определения требований и их предобработки до анализа данных и мониторинга качества. Основные подходы включают создание профиля качества данных с целевыми показателями качества, оценку и улучшение качества на каждом этапе, валидацию и оптимизацию правил качества для повышения точности данных, мониторинг и визуализацию данных после обработки.

С точки зрения принципов FAIR проблемы качества данных должны рассматриваться в контексте обеспечения возможности повторного использования данных (R) [27]. Основной из принципов (R1) в этом направлении указывает на необходимость обильно описывать данные точными и актуальными сведениями (атрибутами). Под такими сведениями подразумеваются ресурсы и метаданные, необходимые для оценки возможности повторного использования данных, а не просто описание их семантики. Эти сведения могут включать информацию о происхождении данных, лицензионные условия их использования и другие нефункциональные свойства, которые тем не менее играют важную роль при выборе данных исследователями для решения своих научных задач [28]. В эту категорию также следует включать метаданные, основанные на показателях качества данных.

К стандартам качества данных, разработанным консорциумом W3C, можно отнести стандарт PROV [29], который устанавливает модель описания про-

исхождения данных. Этот стандарт определяет множество нефункциональных метаданных, которые фиксируются при любой манипуляции с данными. Они включают сведения об авторстве, способе получения, преобразованиях данных, информацию о том, как данные использовались в качестве исходных при создании рассматриваемых данных, и многое другое. Однако на практике доступные данные чаще всего сопровождаются минимальной информацией, ограниченной сведениями об авторстве, принадлежности и времени создания наборов данных, и редко используют все богатые возможности, предоставляемые этим стандартом.

Помимо этого, в рамках консорциума W3C был разработан подход к спецификации качества данных DQV (Data Quality Vocabulary) [30] на основе RDF-технологий. Определения в этой модели позволяют описывать словарь показателей качества наборов данных, включая наборы показателей (dimension) и методы оценки показателей качества (metric), такие как оценки объема, полноты, точности и других характеристик содержимого каталогов данных. Для связи с описываемыми каталогами данных используются спецификации в формате DCAT (Data Catalog Vocabulary) [31].

Спецификация DCAT позволяет описывать каталоги, наборы данных, более общие концепции ресурсов, их представления в различных форматах и источниках, отдельные записи, а также сервисы данных, которые предоставляют доступ к данным через программные интерфейсы для запросов и получения нужных частей данных. Таким образом, спецификации качества данных могут быть связаны как с целыми каталогами и наборами данных, так и с их отдельными фрагментами. Рекомендации DCAT развиваются, в версии 2 добавлены некоторые возможности идентификации данных и обратная связь с моделью DQV, а в недавней версии 3 добавлена поддержка версий и серий наборов данных. Обратная связь с моделью DQV обеспечивает расширение возможностей спецификации качества данных при развитии DCAT.

В целом подходы к построению моделей качества данных можно считать устоявшимися и достаточно гибкими для дальнейшего развития в рамках существующих концепций. Однако, несмотря на это, такие подходы редко применяются для описания данных в виде четко определенных моделей. Метаданные, связанные с качеством данных, чаще всего присутствуют в документации в произвольной форме или включены непосредственно в наборы данных, но не выделены как отдельные метаданные качества и плохо структурированы. Это особенно заметно, например, в случае с погрешностями измерения характеристик объектов. Часто значения характеристик и их погрешности представляются в каталогах как равноценные атрибуты объекта.

3. Метаданные качества и подход к управлению качеством данных

Спецификации качества данных в формате DQV [30] предлагают RDF-схему, которая позволяет задавать структуру для описания качества данных.

Эта схема, среди прочего, включает:

– наборы показателей качества (Dimension): характеристики, которые используются для оценки различных аспектов качества данных;

– категории показателей (Category): классификация показателей по определенным признакам, что облегчает их организацию и использование;

– значения показателей качества (QualityMeasurement): конкретные значения, которые отражают уровень качества по заданным показателям;

– методы оценки показателей качества (Metric): методы вычисления или формирования показателей качества. Они описывают, каким именно способом было оценено качество данных, а также типы данных, которые используются для представления значений показателей качества;

– наборы данных (Dataset) и их представления (Distribution): связывание метаданных качества с конкретными наборами данных или их копиями в определенных форматах;

– сервисы данных (DataService): связывание метаданных качества с сервисами, предоставляющими доступ к данным.

Модель метаданных качества для спецификации качества может быть построена на основе спецификаций, подобных тем, что предлагаются в DQV. Однако такая базовая модель требует значительного расширения, чтобы учитывать спецификацию метаданных качества на различных уровнях агрегации данных, определение действий с данными в зависимости от их качества, установление требований к качеству искомых или создаваемых данных и другие важные особенности.

Для формирования необходимых видов спецификаций, включаемых в модель качества, была проведена классификация метаданных качества по различным критериям. Связывание элементов этой классификации позволяет определить спецификации показателей качества, установить их связь с описываемыми данными, определить действия с данными на основе их качества, а также сформулировать требования к данным, которые необходимо учитывать.

1. Классификация по измерениям показателей качества:

– объем;

– полнота;

– точность;

– происхождение;

– актуальность;

– надежность;

– и др.

2. Классификация по типу значений показателей качества:

– булевый флаг или битовый вектор;

– вещественное значение;

– набор категориальных значений градаций качества;

– набор категориальных значений с различением разновидностей фактов нарушения или выполнения требований качества;

– те же типы с пустыми значениями (неизвестное качество).

3. Классификация по способу спецификации значений:

– константный;

– формулы или правила;

– табличный.

4. Классификация по источнику метаданных:

– метаданные набора данных;

– метаданные в составе схемы данных;

– внешние репозитории метаданных;

– метаданные в публикациях;

– оценка на основании эксперимента;

– статистическая оценка данных;

– оценка на представительной выборке данных;

– изменчивые метаданные (требующие переоценки).

5. Классификация по способу агрегации данных:

– набор данных;

– отношение;

– выборка данных (срез, условия запроса);

– связанная семантика сущностей или процессов;

– кортеж;

– конкретная сущность или процесс;

– атрибут;

– связанная семантика атрибутов;

– значение атрибута.

6. Классификация действий по улучшению качества данных:

– удаление данных в соответствии с уровнем агрегации (набор данных, выборка, кортеж, сущность, атрибут, значение);

– назначение метаданных качества (показатель, вес);

– выбор качественных данных;

– обогащение данных из внешних источников;

– методы усреднения с учетом качества данных;

– методы усреднения без учета качества данных;

– игнорирование качества данных (включение всех данных без учета качества).

7. Классификация требований к качеству данных:

– ограничения на состав и качество исходных данных;

– требования к улучшению качества исходных данных;

– декларируемые необходимые оценки качества результатов;

– характеристика зависимости качества результатов от качества и характеристик исходных данных;

– необходимость оценки качества результатов или невозможности их оценить.

В случаях, когда необходимые метаданные на уровне наборов данных отсутствуют или требуют корректировки на подмножествах данных (например, срезах по параметрам), их можно оценить статистически на всех данных или на их представительных выборках, а затем дополнить существующие метаданные в том же формате.

Метаданные качества на уровне кортежей или значений атрибутов обычно включаются непосредственно в состав данных, представленных в каталогах или наборах данных, в виде дополнительных атрибутов отношений.

Метаданные качества, относящиеся к определенному уровню агрегации данных, распространяются на вложенные уровни агрегации. Таким образом, любой используемый кортеж или значение данных можно связать с метаданными качества, собранными с разных уровней представления данных, включая метаданные набора данных в целом, метаданные атрибутов, значений и др. При формулировании требований к метаданным качества могут использоваться ограничения как на метаданные целых наборов данных, так и на метаданные конкретных значений данных или их происхождения.

Часто применяются подходы к оценке веса качества на основе показателей качества для принятия решений о дальнейших действиях. Вес может выражаться как числовыми значениями, так и специальными обозначениями, такими как отсутствие оценки (nan) или указание на необходимость удаления данных (del). Если значения атрибутов качества отсутствуют, весу качества присваивается значение nan, что не влияет на общий вес качества кортежа. Если хотя бы одно значение веса для кортежа обозначено как del, это означает, что кортеж должен быть исключен из анализа, и другие значения весов не учитываются. Веса качества кортежей, определенные на уровне выборки или набора данных, влияют на общий вес качества кортежа. Наличие нескольких весов для кортежа обобщается путем вычисления их среднего значения. Эти же принципы применяются к весам качества отдельных значений атрибутов. Таким образом, веса качества кортежей и атрибутов также распространяются на веса каждого из значений атрибутов.

4. Пример решения задач с применением спецификаций качества

В качестве примера применения модели управления качеством данных была разработана модель качества, и решалась проблема улучшения качества данных в области звездной астрономии на основе множества больших

многоцветных фотометрических обзоров неба. Для исследования разных типов звезд, их спектров, восстанавливаемых по фотометрическим данным в разных диапазонах излучения, а также их наблюдаемых параметров и оценочных значений астрофизических параметров была поставлена задача перекрестного отождествления и слияния данных об одних и тех же объектах из различных фотометрических каталогов. Пример анализа представлен на данных каталогов SDSS [32], UKIDSS [33], GALEX [34, 35].

Исследование каталогов показывает множество факторов, влияющих на качество данных и отражающихся на качестве отождествления объектов.

– Каталоги имеют разное покрытие неба в связи с расположением обсерваторий, проводивших наблюдения.

– Каждый каталог представлен данными, полученными посредством различных наборов фотометрических фильтров, имеющихся в телескопах и обеспечивающих наблюдения в разных диапазонах спектра излучения объектов.

– Телескопы обладают разным оптическим разрешением и соответственно имеют различную точность позиционирования (калибровки) и различения объектов. Ожидается, что точность различения объектов в наблюдениях в плоскости Галактики может оказаться ниже, поскольку количество объектов в поле наблюдения больше.

– Телескопы имеют разные пределы минимальной и максимальной звездных величин объектов: величины блеска ниже минимального регистрируются в рамках погрешности, а величины выше максимального оказываются засвеченными и не отражают реальные значения.

– Каталоги включают различные наборы флагов качества наблюдений, а также флагов для классификации объектов, обнаружения артефактов и других сведений о наблюдениях.

– Наблюдения, проводимые в разные эпохи, отражаются на позиции объектов вследствие их перемещения и на интенсивности блеска в результате его переменности.

– Искажения данных могут также быть связаны с различными условиями наблюдения: временем года и суток, положением объекта относительно зенита и погодными условиями.

– Расположение объекта на краях области наблюдения влияет на уровень зашумления по сравнению с наблюдениями в центре области, а также может приводить к появлению бликов, связанных с оборудованием.

Задача перекрестного отождествления каталогов и обзоров неба остается актуальной в астрономии и обычно решается для пары каталогов с разными требованиями к качеству отождествления. В соответствии с требованиями конкретных задач проводится предобработка данных или удаление наблюдений с низким качеством. При этом процедура может варьироваться в зависимости от специфики каждого случая.

Исходя из анализа проблемы, были поставлены следующие задачи.

1. Оценка покрытия каталогами: необходимо определить, в достаточной ли степени каталоги охватывают площадки в определенных направлениях обзора неба. Важно, чтобы в одной площадке были данные из нескольких обзоров с разными наборами фильтров, что позволит охватить большинство диапазонов излучения в спектре объектов.

2. Метод перекрестного отождествления: необходимо разработать метод решения задачи перекрестного отождествления обзоров разного качества (разрешения), включающий отождествление множественных наблюдений объектов как внутри одного каталога, так и между различными каталогами.

3. Оценка радиусов отождествления: необходимо решить задачу оценки радиусов отождествления объектов между каталогами в зависимости от направления наблюдения относительно плоскости Галактики.

4. Оценка качества данных: требуется оценить качество данных об объектах наблюдений на основании значений метаданных каталогов и флагов, присутствующих в структурах каталогов.

5. Формирование списков отождествления: необходимо разработать подход к формированию списков отождествления объектов и слиянию данных для получения кортежей, описывающих блеск объектов в разных диапазонах излучения.

6. Представление результатов слияния: требуется представить результаты слияния в виде каталога и связанных с ним метаданных.

4.1. Проблемы полноты данных

Задача 1, связанная с проблемой полноты данных в обзорах, предстает в двух аспектах.

Во-первых, речь идет о покрытии данными используемых каталогов большинства направлений на небе. Это покрытие существенно влияет на решение таких задач, как оценка позиционной точности каталогов и возможность оценки поглощения в межзвездной среде в различных направлениях на небе, что также зависит от угла наблюдения относительно плоскости Галактики. Показатели полноты данных в заданных направлениях оцениваются статистически на основе данных каталогов. В этом контексте важно не процентное покрытие неба, а покрытие крупных площадок в разных направлениях. Для конкретной площадки можно подбирать состав каталогов, который обеспечивает хорошее покрытие.

Во-вторых, полнота данных, представленных фотометрическими каталогами, оценивается с точки зрения покрытия большинства диапазонов наблюдаемого спектра излучения звезд. Эта характеристика имеет решающее значение для оценки спектра на основе данных фотометрии и зависит от набора фильтров, используемых в телескопах. Она определяется набором полей в структуре каталогов.

Очевидно, что данные ни одного из больших фотометрических обзоров в отдельности не соответствуют этим двум требованиям. Именно поэтому

возникает необходимость интеграции и отождествления ряда каталогов, чтобы данные, полученные в результате слияния каталогов, были пригодны для оценки спектра звезд, их дальнейшей классификации и параметризации.

4.2. Проблемы точности данных

В задаче 2 проблема ошибок отождествления множественных наблюдений объектов в каталоге с низким разрешением (GALEX) и затем усреднения ошибочных результатов отождествления до одного кортежа решается с помощью выбора последовательности отождествления каталогов. Сначала все кортежи этого каталога отождествляются с каталогами более высокого качества. Затем, если необходимо получить усредненный кортеж для множественных наблюдений каталога, усредняются результаты перекрестного отождествления с более качественными каталогами. При перекрестном координатном отождествлении объектов из разных каталогов оценивается совместная точность, рассчитываемая как среднеквадратическая ошибка, учитывающая обе неопределенности. Таким образом, достигается меньшая оценка совместной точности (среднеквадратической ошибки) и, соответственно, снижается вероятность неверного отождествления.

При решении задачи 3 радиус отождествления объектов обычно принимается как константа: для каталогов SDSS и UKIDSS – 1 угловая секунда, для GALEX – 3 угловых секунды. Эти значения связаны с позиционной точностью и оптическим разрешением каталогов (точностью различения объектов). Разные значения точности координат можно найти в статьях о каталогах, однако описание качества координат остается неоднозначным. Более того, желательно оценивать радиус отождествления объектов в зависимости от направления на небе. Поэтому обычный подход оказывается недостаточным, и возникает необходимость в определении оптимального радиуса отождествления.

Задача выбора радиуса отождествления решается на основе статистической оценки, которая максимизирует количество уникальных вариантов координатного отождествления для радиуса в конкретном направлении наблюдения относительно плоскости Галактики [36]. Таким образом, по данным или по представительным выборкам определяется точность позиционирования в каталогах, и формируются метаданные спецификаций точности для них. Эти метаданные могут быть представлены таблично в зависимости от площадки в определенном направлении на небе и впоследствии используются для каждого значения координат объектов площадки в качестве его точности.

4.3. Проблемы надежности данных

Задача 4 связана с оценкой точности данных в разных каталогах, которая зависит от целого ряда факторов:

— позиционная точность объектов: зависит от разрешающей способности оборудования, с помощью которого велись наблюдения неба;

— расположение объектов в области наблюдения: объекты, находящиеся в центре области наблюдения, имеют более высокую точность, чем те, что расположены на краю;

— засвеченность фона объектов: яркие звезды, расположенные близко к наблюдаемым объектам, могут создавать помехи;

— артефакты: такие как планеты или спутники, пролетающие в кадре объектов, а также блики от линз;

— засвеченность ярких объектов: яркие звезды на верхней границе звездных величин, определяемых оборудованием, могут искажать данные об их блеске;

— необнаружение объектов: объекты на нижней границе звездных величин могут оставаться незамеченными;

— точность значений блеска: для каждого значения блеска указывается его точность в определенном диапазоне излучения.

Некоторые атрибуты каталогов представляют собой флаги, которые указывают на тип объекта, дополнительные характеристики его наблюдений и их качество. При подготовке данных учитываются ограничения, накладываемые на атрибуты каталогов, не относящиеся к качеству данных, в соответствии с семантикой задачи.

– Атрибуты Fr и Nr в GALEX указывают угловой размер объекта на полувысоте свечения. Ограничением отсекаются кортежи со значениями этих атрибутов, превышающими 0,003 (для звезд).

– Атрибут class в SDSS предоставляет информацию о типе объекта: звезда, галактика, артефакт и др. Ограничением отсекаются кортежи со значениями, отличными от 6 (звезда).

Далее необходимо оценить качество кортежей в целом как наблюдений, так и значений характеристик объектов в них. В каталоге GALEX используются следующие флаги дополнительной информации о наблюдениях и качестве данных.

– Атрибуты Fexf и Nexf – векторы битовых значений, указывающие на различные виды некачественных наблюдений: объекты, имеющие соседей, смешанные с другими объектами, усеченные границей наблюдения и др. При значениях этих флагов, не равных 0, значению блеска присваивается вес качества del.

– Атрибуты Fafl и Nafl – векторы битовых значений, указывающие на различные виды артефактов: блики, пятна и др. При значениях этих флагов, не равных 0, значению блеска также присваивается вес качества del.

– Атрибуты nS/G и fS/G – вещественные значения, оценивающие, является ли объект звездой или галактикой. Можно сказать, что это классификация объекта с указанием степени ее надежности. Назначается вес качества кортежа, равный 1 для значений параметров, превышающих 0,5; в противном случае вес равен 0.

В каталоге SDSS используется один существенный флаг качества кортежа:

– Атрибут Q вводит категориальные градации и оценивает качество наблюдения: 1 – низкое, 2 – приемлемое, 3 – высокое. При значении атрибута, равном 1, назначается вес качества кортежа, равный 0; при значении 2 – вес, равный 0,5; при значении 3 – вес, равный 1.

В каталоге UKIDSS используются следующие флаги качества:

– Атрибут cl – смешанный категориальный и градационный показатель, оценивающий, является ли объект звездой, галактикой или шумом. Значения атрибута могут указывать на отнесение к тому или иному классу с указанием степени надежности. Присваивается значение веса качества кортежа del при всех значениях, кроме -2 и -1 (звезды).

– Атрибут p* – вещественное значение, отчасти дублирующее и уточняющее значение атрибута cl, оценивающее вероятность того, что объект является звездой. При значении атрибута, превышающем 0,7, присваивается значение веса качества кортежа 1; при значении ниже 0,3 – значение 0; иначе значение 0,5.

– Атрибуты pG и pN оценивают подобную вероятность для галактик и шума и не рассматриваются в данной задаче. Кортежи, классифицированные как галактика или шум с помощью атрибута cl, удаляются.

То, каким именно образом формируются веса качества, кортежей или значений на основании значений флагов, было задано постановкой задачи во взаимодействии со специалистами в области звездной астрономии. При отсутствии значений атрибутов качества весу качества присваивается значение nan. Таким образом, на основании флагов были определены веса качества кортежей и некоторых значений атрибутов в кортежах. Обобщение весов качества для получения общих оценок качества кортежей или значений производится в соответствии с принципами, приведенными в предыдущем разделе. Обобщенные веса качества данных учитываются в дальнейшем при слиянии данных.

4.4. Слияние данных

При решении задачи 5, связанной с формированием списков отождествления объектов, применяются принципы работы с отождествляемыми объектами из нескольких каталогов, предложенные в [37]. Используются связки отождествлений, которые исключают транзитивные связи отождествления объектов между несколькими каталогами. Слияние данных производится на основании кортежей, входящих в связки отождествлений, и их обобщенных весов качества.

В зависимости от решаемых задач над отождествленными объектами слияние данных может выполняться по разным принципам. Например, для решения задачи параметризации объектов важны только качественные данные, и любые некачественные наблюдения удаляются из рассмотрения. Однако, если необходимо сохранить некачественные объекты для обеспечения

полноты данных и повышения их качества, такие данные могут быть включены в анализ.

Для идентификации объектов после их слияния и их позиционирования используются координаты из самого точного каталога (в примере это SDSS). Для отождествленных множественных наблюдений одного каталога, собираемых в один кортеж, координаты могут усредняться без учета весов.

Флаги в основном касаются качества наблюдений блесков объектов на небе. Данные блесков при слиянии могут быть усреднены по данным наблюдений в одном диапазоне излучения.

В большинстве каталогов присутствуют поля, указывающие на погрешность (σ) измерения блесков (m). Для оценки погрешности результата слияния при выборе только надежных кортежей (с высоким весом качества: $r = 1$) может использоваться взвешенное среднее значение. При этом вес обратно пропорционален квадрату погрешности, что позволяет минимизировать суммарную дисперсию. Если при решении задачи используются не только надежные кортежи, необходимо учитывать оценку веса качества (r) кортежа или значения совместно с погрешностью значения блеска. В этом случае при вычислении среднего взвешенного (2) вес, обратный квадрату погрешности, умножается на вес качества значения (1).

$$w = \frac{r}{\sigma^2}, \tag{1}$$

$$m = \frac{\sum_i m_i w_i}{\sum_i w_i}. \tag{2}$$

Погрешность результата в этом случае можно оценить следующим образом (3):

$$\sigma = \sqrt{\frac{1}{\sum_i w_i}}. \tag{3}$$

4.5. Формирование результата

Решение задачи 6 заключается в формировании результирующего отношения. Оно создается путем соединения отождествленных кортежей из разных каталогов с обобщенной оценкой атрибутов, для которых были найдены множественные значения. Результат должен объединять в общем кортеже следующие элементы:

— координаты из самого точного каталога, которые также используются для идентификации объекта;

— усредненные взвешенные оценки блесков для каждого диапазона излучения, соответствующего диапазонам каталогов;

— оценки погрешностей для каждого из значений блеска.

Данное отношение может быть представлено в виде самостоятельного каталога, который должен быть обильно снабжен метаданными. Такие метадан-

ные должны включать описание происхождения данных и их качество, что позволит оценивать применимость данных при решении задач и обеспечит возможность их повторного использования.

Описание происхождения данных нового каталога может включать следующие элементы:

— на уровне каталога: ссылки на каталоги, данные которых использовались при формировании результата;

— на уровне атрибутов: ссылки на атрибуты (координаты, блески в соответствующих диапазонах), из данных которых формировались значения атрибутов;

— ссылку на описание метода отождествления кортежей;

— ссылку на описание метода выбора (для координат) и усреднения значений атрибутов (для блесков).

Описание качества данных каталога может включать:

– на уровне отношения информацию о полноте покрытия неба и полноте заполнения атрибутов (охваченных диапазонов наблюдения) в зависимости от направления на небе;

– информацию об оценке позиционной точности результирующих данных;

– информацию о качестве соединенных кортежей;

– информацию о присутствии в атрибутах каталога взвешенных оценок погрешностей блесков и связи их с атрибутами самих значений блеска.

Благодарим О.Ю. Малкова (ИНАСАН) за постановку задачи, используемой для демонстрации представленного подхода.

5. Заключение

Исследованы подходы к спецификации качества данных в исследовательских инфраструктурах с множественными неоднородными источниками данных разного качества. Стандарты описания качества данных опираются на многокритериальные модели, обычно не ограничивающие состав и методы оценки показателей качества. Предложено развитие принципов спецификации качества данных для определения способов хранения и доступа к метаданным качества, уровня агрегации оцениваемых данных и учета нефункциональных требований к данным при решении задач и действий над данными. Эти принципы проиллюстрированы при решении задачи в области звездной астрономии. На разных этапах ее решения оцениваются различные показатели качества, включая полноту, точность и надежность данных, используются разные источники метаданных качества, такие как сведения о точности значений физических величин в каталогах, флаги качества кортежей, гарантированное качество из документации к каталогам, статистические оценки на основе выборки данных каталогов. В зависимости от полученных оценок показателей качества используются разные источники данных, отфильтро-

вываются данные, не подходящие по качеству, изменяется последовательность действий в алгоритмах решения задач, формируются и сохраняются метаданные как источников данных, так и промежуточных и окончательных результатов решения задачи.

СПИСОК ЛИТЕРАТУРЫ

1. *Wand Y., Wang R.* Anchoring data quality dimensions in ontological foundations // Communications of the ACM. New York: ACM, 1996. V. 39. No. 11. P. 86–95.
2. *Ballou D., Pazer H.* Modeling data and process quality in multi-input, multi-output information systems // Management Sci. 1985. V. 31. No. 2. P. 150–162. https://doi.org/10.1287/mnsc.31.2.150
3. *Wang R., Strong D.* Beyond accuracy: What data quality means to data consumers // J. Management Inform. Syst. 1996. V. 12. No. 4. P. 5–33. URL: http://www.jstor.org/stable/40398176
4. *Batini C., Scannapieco M.* Data quality: concepts, methodologies and techniques. Heidelberg: Springer, 2006. 262 p. https://doi.org/10.1007/3-540-33173-5
5. ГОСТ Р 56214-2014. Качество данных. Часть 1. Обзор. М.: Стандартинформ, 2015.
6. ГОСТ Р 57773-2017. Пространственные данные. Качество данных. М.: Стандартинформ, 2017.
7. *Wilkinson M., Dumontier M., Aalbersberg I., et al.* The FAIR Guiding principles for scientific data management and stewardship // Sci. Data 2016. V. 3. Article 160018. https://doi.org/10.1038/sdata.2016.18
8. FAIR data maturity model. Specification and guidelines. Version 1.0. RDA FAIR Data Maturity Model Working Group. Geneva: Zenodo, 2020. https://doi.org/10.15497/rda00050
9. FAIRsFAIR. Fostering FAIR Data Practices in Europe. URL: https://www.fairsfair.eu/
10. *Devaraju A., Mokrane M., Cepinskas L., et al.* From conceptualization to implementation: FAIR Assessment of Research Data Objects // Data Sci. J. 2021. V. 20. No. 1. Article 4. https://doi.org/10.5334/dsj-2021-004
11. The FAIR cookbook for FAIR doers. URL: https://faircookbook.elixir-europe.org/
12. *Harrow J., Drysdale R., Smith A., et al.* ELIXIR: providing a sustainable infrastructure for life science data at European scale // Bioinformatics. Oxford: Oxford University, 2021. V. 37. No. 16. P. 2506–2511. https://doi.org/10.1093/bioinformatics/btab481
13. ELIXIR Platforms. URL: https://elixir-europe.org/platforms
14. Recommendations from the Data Quality Working Group. NASA ES DSWG, 2019. URL: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/recommendations-from-the-data-quality-working-group
15. Data Quality Working Group's comprehensive recommendations for data producers and distributors. NASA ES DSWG, 2019. URL: https://www.earthdata.nasa.gov/s3fs-public/imported/ESDS-RFC-033.pdf
16. ESIP Information Quality Cluster. Earth Science Information Partners (ESIP). URL: http://wiki.esipfed.org/index.php/Information_Quality

17. *Peng G., Privette J., Kearns E., et al.* A unified framework for measuring stewardship practices applied to digital environmental datasets // Data Sci. J. 2015. V. 13. No. 2. P. 231–253. https://doi.org/10.2481/dsj.14-049

18. ISO 19157-1:2023 Geographic information - Data quality. Part 1. General requirements. Geneva: ISO, 2023. URL: https://www.iso.org/standard/78900.html

19. *Sirotnak C., Cook J.* The total economic impact of Talend. Cost savings and business benefits enabled by Talend Solutions. Cambridge: Forrester, 2023. URL: https://www.talend.com/lp/the-total-economic-impact-of-talend/

20. *Chien M., Medd J.* Magic Quadrant for Augmented Data Quality Solutions. Stamford: Gartner, 2024. URL: https://www.gartner.com/en/documents/5257863

21. *Fürber C.* Data quality management with semantic technologies. Thesis. Wiesbaden: Springer Gabler, 2016. https://doi.org/10.1007/978-3-658-12225-6

22. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web // Scientific American 2001. V. 284. No. 5. P. 34–43. URL: https://www.jstor.org/stable/26059207

23. *Cyganiak R., Wood D., Lanthaler M. (eds.).* RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. Wakefield: W3C, 2014. URL: http://www.w3.org/TR/rdf11-primer/

24. *Fürber C., Hepp M.* Towards a vocabulary for data quality management in Semantic Web architectures // Proceedings of the 1st International Workshop on Linked Web Data Management (LWDM2011). New York: ACM, 2011. P. 1–8. https://doi.org/10.1145/1966901.1966903

25. *Hartig O., Zhao J.* Provenance Vocabulary Core Ontology Specification. San Diego: SourceForge, 2012. URL: https://trdf.sourceforge.net/provenance/ns.html

26. *Taleb I., Taleb, Serhani M., Bouhaddioui C., et al.* Big data quality framework: a holistic approach to continuous quality management // J. of Big Data 2021. V. 8. Article 76. https://doi.org/10.1186/s40537-021-00468-0

27. *Gallo R.* Data quality with FAIR principles, an introduction. The Hyve, 2024. URL: https://www.thehyve.nl/articles/data-quality-with-fair-principles

28. *Skvortsov N.* The principles of data reuse in research infrastructures // Proceedings of the International Conference Common Digital Space of Scientific Knowledge: Problems and Solutions (CDSSK 2020). Aachen: CEUR WS, 2021. V. 2990. P. 62–74. URL: https://ceur-ws.org/Vol-2990/rpaper6.pdf

29. PROV-Overview: An overview of the PROV family of documents. W3C Working Group Note. Wakefield: W3C, 2013. URL: http://www.w3.org/TR/prov-overview/

30. Data on the Web Best Practices: Data quality vocabulary. W3C Working Group Note. Wakefield: W3C, 2016. URL: https://www.w3.org/TR/vocab-dqv/

31. *Albertoni R., Isaac A. (eds.).* Data catalog vocabulary (DCAT), Version 3. W3C Recommendation. Wakefield: W3C, 2024. URL: https://www.w3.org/TR/vocab-dcat/

32. *Alam S., Albareti F., Prieto C., et al.* The eleventh and twelfth data releases of the Sloan Digital Sky Survey: Final data from SDSS-III // Astrophys. J. Suppl. Ser. 2015. V. 219. No. 1. P. 12. https://doi.org/10.1088/0067-0049/219/1/12

33. *Lawrence A., Warren S., Almaini O., et al.* The UKIRT Infrared Deep Sky Survey (UKIDSS) // Mon. Not. R. Astron. Soc. 2007. V. 379. No. 4. P. 1599–1617. https://doi.org/10.1111/j.1365-2966.2007.12040.x

34. *Bianchi L., Herald J., Efremova B., et al.* GALEX catalogs of UV sources: statistical properties and sample science applications: hot white dwarfs in the Milky Way // Astrophys. Space Sci. 2011. V. 335. No. 1. P. 161–169. https://doi.org/10.1007/s10509-010-0581-x

35. *Bianchi L., Shiao B., Thilker D.* Revised catalog of GALEX ultraviolet sources. I. The All-Sky Survey: GUVcat_AIS // Astrophys. J. Suppl. Ser. 2017. V. 230. No. 2. P. 24. https://doi.org/10.3847/1538-4365/aa7053

36. *Malkov O., Dluzhnevskaya O., Karpov S., et al.* Cross catalogue matching with Virtual Observatory and parameterization of stars // Open Astronomy 2012. V. 21. No. 3. P. 319–330. https://doi.org/10.1515/astro-2017-0390

37. *Gray J., Szalay A., Budavari T., et al.* Cross-Matching Multiple Spatial Observations and Dealing with Missing Data. Microsoft Technical Report, MSR-TR-2006-175. Redmond: Microsoft Research, 2006. https://doi.org/10.48550/arXiv.cs/0701172

Статья представлена к публикации членом редколлегии А.А. Галяевым.

Поступила в редакцию 29.11.2024

После доработки 10.01.2025

Принята к публикации 14.01.2025