Стандарты и их роль в управлении качеством мастер-данных

Федюшкин Николай Алексеевич
аспирант, программист кафедры автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва
ул. Большевсистска, 68, г. Саранск, 430005, +7-917-992-8356
fedyushkinna@gmail.com

Федосин Сергей Алексеевич
профессор, к.т.н., заведующий кафедрой автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва
ул. Большевсистска, 68, г. Саранск, 430005, (8342) 478691
fedosinsa@mrsu.ru

Савкина Анастасия Васильевна
доцент, к.т.н., доцент кафедры автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва
ул. Большевистская, 68, г. Саранск, 430005, (8342) 290701
av-savkina@yandex.ru

Вечканова Юлия Сергеевна
лаборант кафедры автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва
ул. Большевсистска, 68, г. Саранск, 430005, (8342) 478691
yuliya_kolushova@mail.ru

Аннотация

В статье описываются стандарты качества мастер-данных и их роль в процессе управления мастер-данными. Также рассматривается описание процесса автоматизации управления и руководства данными при помощи методов машинного обучения. Предлагается автоматизация процесса идентификации данных, их классификации и привязки к внутренним и внешним ссылкам для обеспечения семантического значения. Представлена эволюция данных предприятия из «сырого» (необработанного) транзакционного состояния в состояние с высокой степенью маркировки или курируемое состояние. Приведено подробное описание функциональных возможностей процессов управления качеством мастер-данных, которые могут быть решены с помощью методов машинного обучения или других аналитических методов.
The article describes the master data quality standards and their role in the master data management process. We also consider the description of the process of automation of data management and management using machine learning methods. It is proposed to automate the process of data identification, classification and

binding to internal and external links to provide semantic meaning. The evolution of enterprise data from the "raw" (unprocessed) transactional state to the state with a high degree of labeling or supervised state is presented. A detailed description of the functional capabilities of master data quality management processes is given, which can be solved using machine learning methods or other analytical methods.

Ключевые слова

Мастер-данные, стандарты качества данных, метаданные, машинное обучение, автоматизация

Master data, data quality standards, metadata, machine learning, automation

Введение

В последнее время все чаще обсуждаются стандарты, используемые для управления основными данными. Это обусловлено развитием зрелости практики обработки данных, а также требованиями, предъявляемыми к специалистам по обработке больших объемов данных, аналитикам и изменяющимися бизнес-моделями, которые порождают эти две деятельности. С точки зрения передового опыта, стандарты ISO 8000 направлены на решение бизнес-задач, связанных с тем, как точный обмен информацией происходит между деловыми партнерами, особенно в контексте цепочки поставок организации.

В то время как эти стандарты были в разработке в течение некоторого времени, более существенные части были недавно опубликованы в период с 2017 по 2018 годы. В результате видно растущий диалог и интерес со стороны компаний, стремящихся усовершенствовать свои возможности управления основными данными (MDM).

Роль стандартов в управлении качеством основных данных

Как и в случае с любым набором стандартов, важно понимать ценность, которую они приносят организации. Для большинства организаций, имеющих сложные проблемы с цепочкой поставок и проблемы с основными данными, многие используют нынешний подход к исправлению данных вручную, что не приводит к масштабированию больших данных, аналитики и растущих требований к обмену данными. Эти стандарты системно предоставляют возможности, которые имеют значительные преимущества и в конечном итоге снижают накладные расходы, связанные с функцией управления основными данными. Они касаются следующих возможностей:

- классификации, маркировки и систематизированию данных для обеспечения полного раскрытия значения;
- публикации данных в форме, поддерживающей автоматизацию и применение методов машинного обучения;
- оценки данных по качеству (полноте);
- структурированию данных в переносимый формат для возможности совместного использования.

Учитывая это, применение стандартов обходится дорого, и это подробные стандарты, которые для некоторых компаний могут потребовать доработки всего операционного подхода к управлению основными данными (MDM) и руководству данными [1]. Ниже представлены некоторые ключевые факторы и соображения по реализации - как плюсы, так и минусы (табл. 1).

Таблица 1

Ключевые факторы и соображения по реализации стандартов

Применяемый фактор	*Соображения*
Ценность бизнеса	Основные данные по определению являются ценными данными в организации. Как правило, «стоимость» делится на три категории, которые определяют экономическое обоснование. **Операционная эффективность:** плохие данные могут иметь огромное влияние на верхнюю и нижнюю границы. Нехватка запасов будет влиять на доходы и со временем подрывать конкурентные позиции, в то время как производственные проблемы будут снижать маржу, так как ручная корректировка увеличивает затраты. **Отчетность и соответствие:** регулирование в той или иной форме прямо ориентировано на объекты основных данных. Например, защита персональных данных значительно упрощается благодаря возможности использования основных данных опытных клиентов. **Аналитика и понимание:** MDM устанавливает дисциплину в том, как организации определяют термины, решают проблемы; и делает данные более системно пригодными для использования. Все они имеют отношение к тому, как аналитики собирают, систематизируют и применяют к данным методы и модели машинного обучения. Аналитические возможности значительно расширяются, поскольку все метаданные, связанные с классификацией, ссылками связей, семантическими тегами, временными и географическими тегами и анализом образцов, создают богатый набор функций, в отношении которых могут применяться аналитические модели.
Цифровое преобразование	Этот аспект включен только потому, что это то, что стимулирует многие обсуждения данных на исполнительном уровне. Влияние преобразования заключается в том, что текущие бизнес-модели меняются или создаются новые. Многие видят в этом святой Грааль данных и аналитики.
Снижение затрат Операции с данными Качество данных	Движущей силой здесь является автоматизация и обеспечение решения проблемы качества данных до того, как данные попадут в экосистему организации. Это очень важно для организаций, которые решаются на большие объемы данных, поскольку текущие процессы просто слишком громоздки для реализации в реальном времени и не могут масштабироваться для решения проблем с большими данными.
Усложнение/ Упрощение	Этот драйвер может идти в двух направлениях. Эти стандарты часто считаются слишком сложными. Однако внедрение этих стандартов упростит процесс MDM. Организациям необходимо оценить природу их проблемы с основными данными, чтобы понять, перевешивает ли дополнительная сложность данных выгоду для бизнеса.
Отраслевое и партнерское согласование	Эти стандарты касаются того, как можно обмениваться информацией по сложным темам. В некоторых отраслях они уже были приняты или принимаются. Лидерами являются исследования в области здравоохранения/фармацевтики и финансовая отрасль. Часто применение происходит поэтапно. Например, концепция реестра метаданных (ISO 11179) хорошо известна и используется во многих отраслях. Спецификации ISO 8000, однако, являются более новыми и, следовательно, менее реализованными.

Рекомендация по внедрению стандартов в организации состоит в том, чтобы понять основные концепции, связанные со стандартами, и включить эти возможности в стратегию данных и план действий. Например, нельзя сделать данные «переносимыми», можно добавить детали в словарь для поддержки процесса. *Используя этот подход, можно постепенно построить более зрелый набор возможностей, которые упрощают и оптимизируют процесс MDM во всей цепочке поставок* [2].

Стандарты качества данных MDM

Рассматриваем ISO 8000 как общий стандарт, поскольку он неразрывно связан с двумя другими стандартами: 1) Стандарты описания данных (ISO 22745) и 2) как информация о данных должна быть представлена партнерам по обмену данными (ISO 11179). Эти стандарты действительно учитывают то, как организации управляют информацией о своих метаданных, другими словами, метаданными о метаданных.

Это всеобъемлющие стандарты, каждый из которых имеет много глав (в ISO - «детали/части»). Приведенная ниже информация фокусируется на ключевых элементах, которые влияют на обсуждение цепочки поставок. Экономическое обоснование для каждой организации определяет, какие концепции принимаются, и в какой степени реализуются все стандарты [3].

ISO 8000: Метаданные о словаре данных

Стандарт ISO 8000 позволяет организациям обмениваться данными, зная, что они обмениваются данными высокого качества. В частности, основное внимание уделяется основным данным цепочки поставок - хотя нет причин, по которым стандарт не следует применять более широко.

Стандарт ориентирован на то, как организации:

- системно упаковывают и передают информацию о своих данных - делая данные переносимыми;
- создают метаданные, которые позволяют оценить соответствие с набором спецификаций;
- убеждаются, что данные помечены полностью и правильно в отношении: синтаксиса, семантического кодирования, соответствия требованиям, происхождения и точности.

ISO 22745: Словарь данных, описанный в ISO 8000

Этот стандарт описывает, из чего состоит информация, которая должна быть в словаре данных, чтобы иметь совместно используемые данные. Эти словари называются «открытыми техническими словарями», так как предполагается, что они будут использоваться или доступны для всех участников цепочки поставок.

Важнейшим аспектом этого стандарта являются его требования к использованию уникальных идентификаторов концептов для создания однозначных, не зависящих от языка описаний отдельных лиц, организаций, мест, товаров, услуг, процессов, правил и положений.

Если в качестве примера использовать среду розничной торговли, «Идентификаторы концепций» будут определять связанные словари, организованные вокруг категорий розничных продуктов. Так, чтобы описать товары, которые являются

предметами одежды, можно было бы обозначить их как «летняя одежда» или «зимняя одежда»; также добавить мужская это одежда или женская. Организация словарей таким образом известны как онтологии.

Процесс использования идентификаторов концептов из внешнего открытого технического словаря является формой семантического кодирования, совместимого с ISO 8000.

ISO 11179: Реестр, в котором содержатся идентификаторы концепций, необходимые в словаре

Этот стандарт имеет более широкий охват, чем просто держатель идентификаторов концептов. Можно рассматривать Реестр как хранилище метаданных, которое содержит связанные словари, используемые для семантического описания данных в Словаре данных. Запись словаря данных связана с термином в Реестре, так что запись может быть полностью описана «в контексте». Например, менеджер цепочки поставок для розничного магазина может иметь категорию продукта «Церемонный Китай» (Formal China). В этом случае «Китай» будет связан с понятием «Посуда». Если розничный магазин также занимался одеждой, там также может быть запись «Церемонный Китай», в которой упоминается одежда, используемая китайцами для официальных мероприятий. В этом случае «Китай» будет связан с концептуальным термином в Реестре, который называется «Страна». Используя этот подход, логика может быть закодирована в программное обеспечение управления данными для выполнения правил основных данных, включающих семантические знания.

Ключевой концепцией 11179 является то, что реестр, в котором хранятся данные, открыт и может быть доступен третьим сторонам. Реестры могут вестись и публиковаться покупателем, продавцом или третьей стороной (как правило, отраслевой ассоциацией или руководящим органом).

На первый взгляд, все эти идеи имеют смысл. Проблема заключается в том, что данные, которые помечены или курируются до этого уровня, часто требуют значительного обновления возможностей. Классификация данных, управление иерархически связанными словарями данных и дисциплина управления/руководства для выполнения всей этой работы, вероятно, потребуют эволюции в зрелости управления данными.

Для многих способность создавать классификацию данных и возможности управления словарем представляет собой самую большую неизвестность. Для упрощения и автоматизации управления качеством данных в цепочке поставок рассмотрим возрастающую роль машинного обучения

Автоматизация управления и руководства данными с помощью машинного обучения

Автоматизируем процесс идентификации данных, их классификации и привязки к внутренним и внешним ссылкам для обеспечения семантического значения [4] для описания машинного обучения менеджера данных и выполняемых им задач в контексте основанного на стандартах операционного контекста.

Рассмотрим эволюцию данных из «сырого» (необработанного) транзакционного состояния в состояние с высокой степенью маркировки или курируемое состояние, которое может быть разделено между покупателем и продавцом (рис. 1). Символы базы данных, выделенные синим цветом (сплошные

линии), представляют данные в состоянии покоя. Прямоугольные элементы, выделенные зеленым цветом (пунктирные линии), представляют задачи, которые автоматизируют процесс увеличения данных при их перемещении по этому пути.

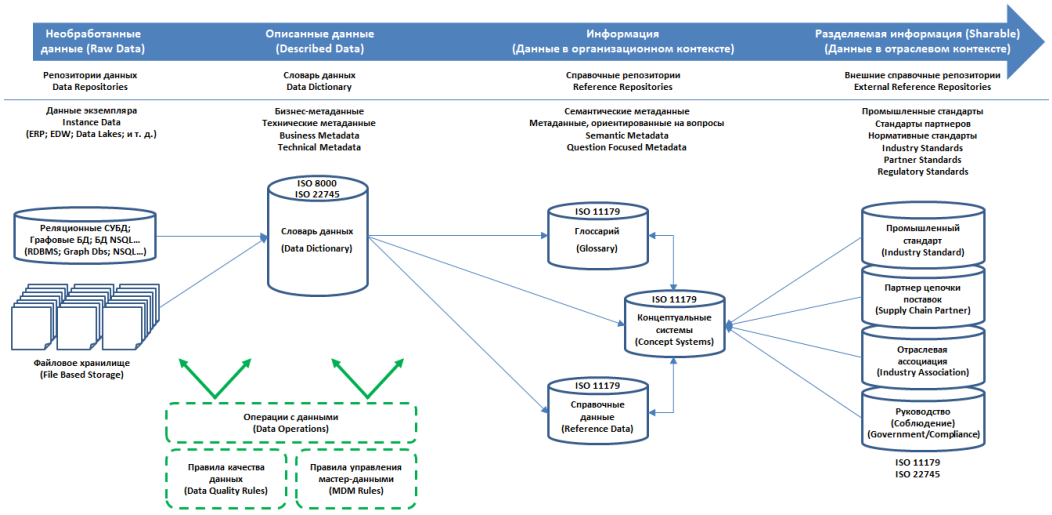


Рис. 1. Курирование от необработанной информации к разделяемой информации (с совместным доступом)

Действия в рамках задач «Правила качества данных» и «Правила MDM» можно разбить на ряд функциональных возможностей. Некоторые из этих возможностей являются традиционными задачами обработки данных, а именно, сохранение или удержание метаданных в базе данных и предоставление данных с помощью некоторой возможности каталогизации и публикации. Другие элементы (обведены синим цветом) - это те, где могут применяться подходы машинного обучения (рис. 2).

Рис. 2. Функциональные возможности в рамках задач «Правила качества данных» и «Правила управления MDM»

Машинное обучение имеет несколько определений в популярной литературе, среди них исчерпывающее определение сайта Techemergence:

«Машинное обучение - это наука о том, как заставить компьютеры учиться и действовать так, как это делают люди, и совершенствовать свое обучение с течением времени автономно, предоставляя им данные и информацию в форме наблюдений и взаимодействия в реальном мире».

Методы машинного обучения играют важную роль в автоматизации описанного выше процесса, особенно в отношении неизвестных или новых наборов данных.

Для специалистов по управлению данными важно понимать, что ни одна техника машинного обучения не будет применяться единолично. По всей вероятности, несколько подходов будут объединены воедино и часто выполняются рекурсивно, чтобы гарантировать, что данные могут быть идентифицированы, классифицированы, а затем связаны с соответствующим уникальным идентификатором. В идеальном мире

алгоритмы будут меняться или учиться приспосабливаться к изменениям классифицируемых данных [5]. В таблице 2 перечислены некоторые методы машинного обучения, которые могут быть применены.

Таблица 2

Методы машинного обучения

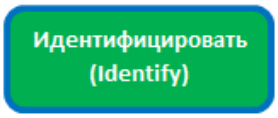
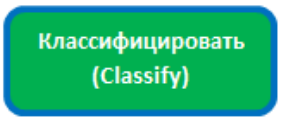
Методы машинного обучения	
Неструктурированные данные	*Структурированные данные*
Связывание сущностей / Извлечение Категоризация Кластеризация Суммаризация Тегинг (пометка) Связывание	Ассоциирование Характеризация Классификация Прогнозирование Кластеризация Исследование образов Анализ исключений

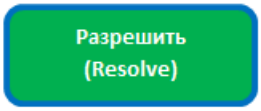
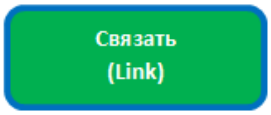
Стоит обратить внимание, что они неизменно взаимодействуют друг с другом. Если помечать сущности людей в неструктурированном тексте, можно пожелать охарактеризовать их, используя структурированную технику: подсчет мужских имен, частоту в текущем документе, частоту во всех документах и т. д. Это говорит о многоуровневом и рекурсивном характере машинного обучения и богатстве метаданных, которыми должна управлять команда по работе с этими данными [6,7].

Они подробно описаны в таблице 3 с соображениями для руководителей программ.

Таблица 3

Подробное описание функциональных возможностей, которые могут быть решены с помощью методов машинного обучения или других аналитических методов

Возможность	*Соображения*
Идентифицировать 	Подходы машинного обучения поддерживают идентификацию данных экземпляра для классификации данных. Для организаций, где существует значительная установленная устаревшая проблема, будет важно иметь алгоритмы, которые идентифицируют данные, представляющие интерес. Идентификация личной информации - актуальная область интересов, обусловленная регулированием GDPR (общий регламент по защите персональных данных).
Классифицировать 	Как только данные идентифицированы, подходы машинного обучения поддерживают классификацию данных в словаре данных: данные находятся в финансовой области; находится в фазе «Доставка» жизненного цикла Справочника по операциям с цепочками поставок (SCOR - Supply Chain Operations Reference). Должны существовать алгоритмы классификации, которые помечают данные соответствующим классификатором. Возможности должны количественно определять и разрешать те случаи, когда существует неопределенность, чтобы сделать алгоритм классификации более точным.
Разрешить	Законченный словарь данных будет поддерживать разрешение объектов, предоставляя более богатый набор функций, в отношении которых можно запускать алгоритмы машинного обучения MDM. Для восстановления идентичности элемента основных данных может потребоваться итеративный запуск

	многоуровневого подхода: применяется Алгоритм #1; для тех, которые не разрешаются с помощью алгоритма #1, применяется алгоритм #2; и т. д.
Связать 	Разрешенный объект должен быть связан с внутренними и внешними справочными источниками. Методы машинного обучения могут использоваться для идентификации и определения кандидатов на ссылки и определения типа/силы ссылки. Аналитические детали этого подхода могут быть рассмотрены в вышеупомянутой возможности «Разрешить». Тем не менее, в центре внимания здесь должно быть определение правильной ссылки (или ссылок), где есть несколько возможных наборов ссылок, где могут быть установлены ссылки. Это критический шаг, так как связь с внутренней ссылкой «Концептуальная система» - это то, что описывает данные элемента с семантической точки зрения. Это также то, что связывает описываемые данные с общедоступным набором определений, на которые могут ссылаться внешние стороны (рис. 1). Эти связи пересекают общепринятое в отрасли определение между партнерами по цепочке поставок. Пример: если менеджер по цепочке поставок пытается сообщить о характере требований к продукту поставщика, например, крепежному винту, возможность указать длину винта в зависимости от длины «плеча» на винте; размер резьбы (метрическая, стандартная, империческая); тип головки (шестигранник, квадрат, панорамирование и т. д.) имеет решающее значение. Внутренние метки для них связаны с отраслевыми согласованными метками, доступными для сообщества поставщиков. Пока поставщик использует одну и ту же систему эталонных концепций, и покупатель, и продавец могут быть уверены, что они говорят об одном и том же крепежном винте.

После завершения этих действий результаты необходимо сохранить в хранилище метаданных и опубликовать в каталоге данных, который позволит пользователям понять, какие данные доступны и как к ним можно получить доступ.

Заключение

Проделанные исследования очень важны для понимания студентами и магистрантами того, как стандарты и машинное обучение вписываются в информационную архитектуру, и могут использоваться в качестве дополнительного материала по дисциплинам «Алгоритмы и методы машинного обучения», «Организация и управление предприятием», «Системы хранения данных», при прохождении производственной практики в компаниях, стремящихся усовершенствовать свои возможности по управлению основными данными (MDM), с учетом стандартов. В качестве вывода можно отметить, что для организаций с устоявшейся и зрелой функцией управления многие из рассмотренных вопросов будут успешно решены, тогда как для организаций, которые имеют меньшую зрелость возможностей, стратегия и пути их развития должны быть явными при определении

бизнес-единиц, в которых могут быть созданы фундаментальные возможностях по мере развития потребностей и зрелости.

Литература

1. Jonathan Adams. Role of Standards in Managing Master Data Quality // The Data Administration Newsletter. 05.09.18. URL: http://tdan.com/role-of-standards-in-managing-master-data-quality/23707 (дата обращения: 23.04.19).
2. Pierre Bonnet. Enterprise Data Governance: Reference and Master Data Management Semantic Modeling. // ISBN: 978-1-118-62253-7., Wiley-ISTE, 2013, 320 pages
3. Jonathan Adams. MDM Data Quality Standards // The Data Administration Newsletter. 03.10.18. URL: http://tdan.com/mdm-data-quality-standards/23828 (дата обращения: 23.04.19).
4. Федюшкин Н.А., Федосин С.А., Савинов И.А., Латентно-семантический анализ текста // Актуальные проблемы технических наук в России и за рубежом. Сборник научных трудов по итогам международной научно-практической конференции — № 5 — г. Новосибирск — 2018 — С. 15-17
5. Jonathan Adams. Automating Data Management and Governance through Machine Learning // The Data Administration Newsletter. 07.11.18. URL: http://tdan.com/automating-data-management-and-governance-through-machine-learning/23972 (дата обращения: 23.04.19).
6. Федюшкин Н.А., Федосин С.А., Основные технологии интеллектуального анализа текста // Развитие технических наук в современном мире. Сборник научных трудов по итогам международной научно-практической конференции — № 3 — г. Воронеж — 2016 — С. 21-25
7. Федюшкин Н.А., Федосин С.А., Краткий обзор методов и моделей интеллектуального анализа текста // Проблемы и достижения в науке и технике. Сборник научных трудов по итогам международной научно-практической конференции — № 4 — г. Омск — 2017 — С. 10-12