Метаописание и верификация классификаторов в автоматизированной системе

А. С. Чайковская, В. В. Баранюк

***Аннотация* – Статья посвящена вопросам метаописания классификаторов в автоматизированной системе (АС) и проверки корректности их содержимого. Система метаданных является источником информации о содержимом ресурса и основой для реализации функций поиска ресурсов, управления ими в системе и обмена с другими системами. Применение общепринятых форматов метаописаний для описания классификаторов нецелесообразно ввиду их универсальности – в них входит много неиспользуемых атрибутов и отсутствуют необходимые. В статье приводится формат метаописания, разработанный на основе стандарта Дублинское ядро, учитывающий специфические особенности классификаторов как информационных ресурсов, их жизненного цикла. Затем выполняется анализ возможных аспектов и предлагается подход к автоматизированной верификации классификаторов в АС, который дает возможность обнаруживать ошибки различных классов, приводящие к искажению данных и другим негативным последствиям, вплоть до формирования некорректных решений задач автоматизированной системой. Это, в свою очередь, может стать причиной экономического, репутационного и других видов ущерба для организации, применяющей эти решения в своей деятельности. Дальнейшая корректировка найденных ошибок позволяет существенно повысить качество информационного обеспечения АС, что, в свою очередь, положительно сказывается на качестве решения задач в АС.**

Ключевые слова – автоматизированные системы, информационные ресурсы, классификаторы, метаданные, ведение классификаторов, метаописание классификаторов, верификация классификаторов.

I. Введение

Век информационных технологий отразился на всех сферах жизни человека. В настоящее время происходит повсеместная автоматизация и многократное усиление взаимодействия между различными ведомствами. В Государственной программе «Информационное общество 2011 – 2020» обозначены задачи создания электронного правительства и повышения эффективности государственного управления, которые в частности предусматривают:

– создание и развитие государственных межведомственных информационных систем, предназначенных для принятия решений в реальном времени;

– создание справочников и классификаторов, используемых в государственных и муниципальных информационных системах;

– обеспечение информационной поддержки руководителей;

– сокращение временных и финансовых затрат, вызванных несовместимостью информационно-телекоммуникационных систем, дублированием подготовки данных, их противоречивостью, затруднениями с доступом, выборкой и передачей информации и др. [1].

Для корректной, слаженной работы этих систем необходима их совместимость на различных уровнях, в частности, на уровне информационного обеспечения. Оно включает средства описания данных – классификаторы – и типы данных, включая их наименования, идентификаторы и атрибуты. Метаданные определяют набор характеристик данных таким образом, чтобы их можно было интерпретировать самостоятельно или рассматривать эти данные как определенную информацию.

Информационная совместимость – это способность двух или более систем адекватно воспринимать одинаково представленные денные. Обеспечение информационной совместимости АС дает существенный экономический эффект в первую очередь за счет резкого сокращения дублирования в описании данных и устранения несопоставимости описаний одних и тех же данных. Кроме того, благодаря установлению методического и технологического порядка при описании данных в автоматизированных системах, удается существенно снизить финансовые затраты на разработку информационного обеспечения АС.

Система метаданных необходима для управления информационными ресурсами в АС. Несмотря на то, что большое количество систем успешно используют различные форматы работы с метаданными, классификаторы как информационные ресурсы имеют свою специфику, что сказывается на составе метаописания. Одной из целей данной статьи как раз и является формирование предложений по составу метаданных классификаторов в АС, предоставляющих только необходимую и достаточную информацию.

Статья получена 9.05.2018 г.
А. С. Чайковская, НИЯУ «МИФИ» (e-mail: chaikovskaia.a@gmail.com).
К.т.н., с.н.с., В. В. Баранюк, МИРЭА (e-mail: valentina_bar@mail.ru).

Качество решения задач в автоматизированных системах напрямую зависит от качества используемого информационного обеспечения. Однако даже принятые на государственном уровне классификаторы, вследствие специфики процесса их разработки и ведения, зачастую содержат значительное число ошибок разного рода, начиная от некорректных кодов классификационных единиц и заканчивая ссылками на несуществующие позиции. Таким образом, важной проблемой является верификация классификаторов. В данной статье предлагается подход к автоматизированной проверке корректности классификаторов, на основе которого возможна разработка методики верификации классификаторов в АС.

II. Метаописание классификаторов

Согласно [2], под классификатором понимается информационный ресурс, распределяющий информацию в соответствии с ее классификацией (классами, группами, видами и другими признаками).

Источник [3] определяет классификатор технико-экономической и социальной информации как нормативный документ, устанавливающий систематизированный перечень наименований и кодов объектов классификации и/или классификационных группировок и принятый на соответствующем уровне стандартизации.

По своему статусу классификаторы являются нормативными документами по стандартизации, которые разрабатываются по определенным правилам, утверждаются (принимаются) в установленном порядке и являются обязательными для применения в соответствующих сферах управления и областях деятельности.

Метаданные определяют набор характеристик данных таким образом, чтобы их можно было интерпретировать самостоятельно или рассматривать эти данные как определенную информацию. Согласно [4], метаданные – это структурированные данные, характеризующие информационный ресурс для целей его идентификации, поиска и управления им.

Принципиально важным является то, что системы метаописания должны решать задачи формализации и стандартизации записей метаданных и обеспечения возможности автоматизированной обработки метаописаний.

Дýблинское ядро (англ. DublinCore, DC) — стандарт метаданных (формат метаданных), простой и эффективный набор для описания широчайшего диапазона сетевых ресурсов [4]. Это один из самых успешных проектов, связанных с описанием ресурсов, он широко используется при разработке крупных систем, работающих с метаописаниями. В основе Дублинского ядра лежит разработка структуры метаописаний ресурсов.

Основные элементы метаданных Дублинского ядра разбиваются на 3 группы, которые соответствуют классу или области информации, хранящейся в них:

– элементы, относящиеся к описанию содержания ресурсов;

– элементы, относящиеся к интеллектуальной собственности;

– элементы, относящиеся к идентификации ресурсов.

Элементы DC в настоящий момент входят практически во все системы метаданных.

Структурно Дублинское ядро состоит из двух уровней, используемых в зависимости от необходимости детализации описания информационного ресурса:

– SimpleDublinCore – простой (неквалифицированный) уровень;

– QualifiedDublinCore – компетентный (квалифицированный) уровень.

По названию уровней иногда называют модели Дублинского ядра, например, простая модель.

На первом уровне простой набор элементов метаданных Дублинского ядра состоит из 15 атрибутов метаданных:

Title — название;

Creator — создатель;

Subject — тема;

Description — описание;

Publisher — издатель;

Contributor — сущность, участвовавшая в создании ресурса (организация, сервис, человек);

Date — дата;

Type — тип;

Format — формат документа;

Identifier — идентификатор;

Source — источник;

Language — язык;

Relation — отношения;

Coverage — покрытие;

Rights — авторские права.

На втором уровне квалифицированный (компетентный) набор элементов метаданных Дублинского ядра, помимо 15 вышеперечисленных, может включать:

Audience — аудитория (зрители);

Provenance — происхождение;

RightsHolder — правообладатель.

Атрибуты, входящие в состав даже простого уровня Дублинского ядра, не полностью применимы к описанию классификаторов в АС. В то же время, большую часть характеристик классификаторов невозможно отобразить с помощью Дублинского ядра. Однако этот стандарт дает общий подход и начальные данные для разработки собственной системы метаописания.

Специфические особенности классификаторов могут быть определены в первую очередь на основании документа [5]. Предлагаемый перечень атрибутов метаданных классификаторов приведен в таблице 1.

Атрибуты Title, Description и Identifier были заимствованы из стандарта «Дублинское ядро» без изменений. Элементы, содержащие информацию об авторстве и проч., были переработаны в соответствии с [2]. На основании этого же документа в метаописание

включены код по ОКОК, дата утверждения и дата введения классификатора. Также введены атрибуты, относящиеся к внутренней структуре классификатора, описывающие используемые методы классификации и кодирования. В метаописании указывается категория классификатора. Также добавлены атрибуты для ведения регистрационных данных классификатора по предприятию, в рамках которого функционирует АС – это нормативный документ, на основании которого организация получает классификатор, а также регистрационный номер и идентификатор в журнале учета информационных ресурсов организации.

Таблица 1 – Система метаданных классификаторов для АС

Атрибут	Название атрибута	Описание
Title	Наименование	Наименование классификатора
ShortTitle	Краткое наименование	Краткое наименование (аббревиатура) классификатора
UID	УИД	Уникальный идентификатор ресурса в системе
Description	Описание	Словесное описание классификатора
ReceiptGround	Основание, дата, №	Нормативный документ, на основании которого организации поставляется данный классификатор
RegistrationNumber	Регистрационный номер	Номер записи, соответствующей классификатору, в журнале учета информационных ресурсов организации
Identifier	Идентификатор	Внутренний идентификатор классификатора по регистрационному журналу предприятия
ApprovalDate	Дата утверждения	Дата утверждения классификатора ответственной организацией
ImplementationDate	Дата введения	Дата введения классификатора в действие
TableName	Имя таблицы	Наименование таблицы содержимого классификатора в базе данных
Developer	Организация-разработчик	Организация, ответственная за разработку классификатора
Maintainer	Ведущая организация	Организация, ответственная за ведение классификатора
Affirmant	Утверждающая организация	Организация, утверждающая классификатор
ClassificationMethod	Метод классификации	Метод классификации, по которому построен классификатор
CodingMethod	Метод кодирования	Метод кодирования, в соответствии с которым кодируются классификационные единицы
Category	Категория	Категория, к которой относится классификатор
CodeOK	Код по ОКОК	Код по Общероссийскому классификатору общероссийских классификаторов (только для классификаторов из категории общероссийских)
UpdatePeriodicity	Периодичность обновления (мес.)	Периодичность, с которой должно производиться обновление содержимого классификатора

III. Верификация классификаторов

Порядок разработки общероссийских классификаторов (ОК) устанавливается в [2]. Основные положения этого стандарта применяются и при создании классификаторов других категорий.

В соответствии с этим стандартом общероссийские классификаторы включают в себя следующие структурные элементы:

- обложку;
- титульный лист;
- предисловие;
- содержание;
- наименование общероссийского классификатора;
- дату введения;
- введение;
- перечень позиций;
- приложение.

Следует обратить внимание на то, что, говоря о верификации классификатора, авторы имеют ввиду его перечень позиций.

При разработке первой редакции проекта классификатора проводятся следующие работы:

- классификация заданного множества объектов классификации;
- унификация построения и написания наименований объектов классификации;
- кодирование заданного множества объектов классификации.

Процесс классификации очень сложный и творческий, требует широких познаний в рассматриваемой области, поэтому автоматизация проверки правильности классификации не представляется возможной. Унификация наименований объектов также остается за оператором. А вот некоторые аспекты проверки кода могут быть в значительной мере автоматизированы.

Кодирование объектов классификации предусматривает:

- выбор метода кодирования;
- выбор алфавита и длины кода;
- построение структуры кода;
- кодирование объектов классификации и их группировок;
- расчет, при необходимости, контрольных чисел для защиты кодов классификатора;
- обеспечение резервной емкости кодов классификатора.

Раскроем сущность этих пунктов с помощью определений из [2].

Код — это знак (символ) или совокупность знаков (символов), принятых для однозначного обозначения классификационной группировки или объекта классификации. Таким образом, обязательно должна проводиться проверка **наличия кода** (в общем случае должно быть заполнено хотя бы одно из кодовых полей), а также исключено **дублирование кодовых обозначений.**

Основными методами кодирования объектов технико-экономической и социальной информации являются последовательный, параллельный, порядковый и серийно-порядковый [5]. Эти методы состоят в следующем [6]:

- последовательный: в кодовом обозначении по очереди указываются зависимые признаки классификации;
- параллельный: в кодовом обозначении указываются независимые признаки классификации;
- порядковый: кодовыми обозначениями служат числа натурального ряда;
- серийно-порядковый: кодовыми обозначениями служат числа натурального ряда с закреплением отдельных диапазонов (серий) за классами кодируемых объектов.

Применяемый метод кодирования существенно сказывается на структуре кода.

Алфавит кода – система знаков (символов), принятых для образования кода. Так как набор допустимых символов фиксирован, может быть реализована проверка кодов на **наличие символов, не входящих в алфавит кода.**

Число знаков в коде называют длиной кода. Все объекты классификатора должны иметь **коды заданной длины**. Это еще один аспект верификации кодов. Данный пункт может быть совмещен с общей для всех полей проверкой на длину значения.

Структура кода определяется исходя из трех предыдущих пунктов. Как правило, на определенных позициях в коде могут находиться только знаки из определенного множества. То есть отдельные участки кода могут иметь свой алфавит. Это позволяет реализовать **проверку по маске кода.**

Некоторые классификаторы содержат контрольные числа для кодовых полей. Контрольное число может находиться как в составе кода, так и в отдельном поле таблицы классификатора. **Верификация кода по контрольному числу** также может быть автоматизирована. От оператора в данном случае требуется задать местонахождение исходных данных для расчета контрольного числа, используемую методику расчета и расположение эталонного контрольного числа. Как правило, используются стандартные методики расчета, вследствие чего они могут быть заранее реализованы в АС.

При необходимости, резервная емкость классификатора также может вычисляться средствами АС, однако к верификации классификаторов она не имеет прямого отношения.

Помимо кода, наименования и, возможно, контрольного числа, каждая позиция классификатора может также включать дополнительные классификационные признаки [2], если они предусмотрены. В качестве дополнительных признаков могут быть использованы коды взаимосвязанных классификаторов. В некоторых общероссийских классификаторах, имеющих связи с другими,

обнаруживаются ссылки на несуществующие коды взаимосвязанных классификаторов. Автоматизированная **проверка классификатора по связям** позволяет избежать этих проблем.

Дополнительные признаки также могут иметь специальный тип. Одним из них является УИД позиции классификатора. Необходима **валидация УИДа на соответствие установленному формату**, а также проверка на **дублирование по УИДу**. Хотя при автоматической генерации УИДа вероятность совпадения крайне мала, возможен некорректный ручной ввод значений.

Другой специальный тип – статус позиции. **Поле «статус» может иметь 4 возможных значения**:

– 0 – первоначальная загрузка;
– 1 – позиция удалена;
– 2 – позиция изменена;
– 3 – позиция добавлена.

Полю статус обычно сопутствуют дата утверждения и дата введения изменения. Вся эта информация может использоваться в качестве основы для **контроля корректности последовательности изменений** позиций классификатора. Так, позиция не может быть изменена до того, как была добавлена, и не может быть удалена, если она отсутствовала ранее.

Во всех остальных случаях могут использоваться общие методы контроля содержимого исходя из известных **типа и длины значения признака**, заданной **точности числа**, если это значение вещественное, а также **требований к заполненности значения**. Кроме того, могут пригодиться **условия проверки на наличие символов из заданного набора**. Например, иногда возникают ситуации, когда в тексте на кириллице встречаются латинские буквы, аналогичные по написанию, но отличные в контексте машинной обработки. Подобные проверки позволяют обнаруживать определенную долю опечаток в значениях признаков.

Также может быть реализована **проверка орфографии.** Так как тип полей классификатора заранее известен, проверка орфографии может распространяться только на текстовые поля.

В завершение рассмотрим **проверку на наличие нечитаемых и запрещенных символов**. К ним относится большинство управляющих символов, за исключением символа возврата каретки и переноса строки. Они не несут смысловой нагрузки, зато могут порождать ошибки при машинной обработке данных и влиять на их представление. Например, null-символ в строке распознается как конец строки в ряде систем, что может приводить к их уязвимостям или некорректной работе [7]. Также запрещена табуляция – внутри перечня позиций классификатора она не используется.

Все выделенные в данном разделе аспекты верификации классификаторов могут быть учтены при создании соответствующей методики для использования в автоматизированной системе. Обобщая вышесказанное, можно представить процесс верификации классификаторов в следующем виде:

а) проверка позиций классификатора:
1) проверка на наличие нечитаемых и запрещенных символов;
2) проверка кодовых полей:
– заполненность хотя бы одного кодового поля;
– проверка отсутствия знаков, не входящих в алфавит кода;
– валидация длины кода;
– соответствие кода маске кода;
– совпадение контрольного числа;
3) наличие значений в обязательных полях;
4) проверка типов значений признаков;
5) валидация длины значений признаков или точности числа (для вещественных значений);
6) корректность значений признаков по дополнительным условиям проверки;
7) соответствие значений в полях «УИД» общепринятому формату;
8) проверка допустимости значения поля «статус»;
9) корректность последовательности изменений записи по полю «статус»;
10) проверка связей с позициями других классификаторов;
11) проверка орфографии;
б) проверка на дублирование записей по кодовым полям;
в) проверка на дублирование записей по УИДу.

IV. Выводы

В данной статье предложена специализированная система метаданных, которая представляет собой неизбыточное описание классификаторов как информационных ресурсов. Данное метаописание содержит информацию, позволяющую не только реализовать автоматизированный поиск и управление классификаторами в АС, но и дающую представление о содержимом и структуре классификатора, что упрощает работу с ним.

Также на основе анализа структуры классификаторов, процессов их разработки и ведения выявлены аспекты верификации классификаторов. Последовательная их реализация в процессах разработки и ведения классификаторов в АС позволит значительно повысить качество классификаторов, исключив большую часть ошибок, что, в свою очередь, отразится на качестве решения задач в АС.

Из-за наличия в тексте на русском языке аналогичных по начертанию русским буквам символов латиницы у пользователей возникают трудности при поиске информации – такие записи просто не будут найдены.

Дублирование записей по кодовым полям эквивалентно неоднозначности информации, представляемой классификатором, хотя одна из ключевых его ролей состоит как раз в обратном. Это или провоцирует сбои в функционировании АС, или приводит к неверным решениям при решении задач в АС. Например, может быть выбран не тот пункт назначения для доставки груза.

Таким образом, применение предложенных этапов верификации классификаторов позволит избежать множества проблем при их разработке, ведении и использовании.

БИБЛИОГРАФИЯ

[1] Распоряжение Правительства Российской Федерации от 20 октября 2010 г. № 1815-р «О государственной программе Российской Федерации "Информационное общество (2011 – 2020 годы)"».

[2] ПР 50.1.024–2005. Основные положения и порядок проведения работ по разработке, ведению и применению общероссийских классификаторов. – М.: Стандартинформ, 2006. – 34 с.

[3] ГОСТ 1.1–2002. Межгосударственная система стандартизации. Термины и определения. – М.: ИПК Изд-во стандартов, 2002. – 30 с.

[4] ГОСТ Р ИСО 15836-2011. Информация и документация. Набор элементов метаданных DublinCore. – М.: ФГУП «Стандартинформ», 2014 – 12 с.

[5] ПР 50.1.024-2005. Основные положения и порядок проведения работ по разработке, ведению и применению общероссийских классификаторов. — Введ. 01.04.2006. — М.: Стандартинформ, 2006. — 34 с.

[6] Мамиконов А.Г. Основы построения АСУ. — М.: Высшая школа, 1981. — 248 с.

[7] Null-terminated string [Электронный ресурс] / Электрон. Текстовые дан. – Режим доступа: https://en.wikipedia.org/wiki/Null-terminated_string, свободный.

Classifiers' meta-description and verification in the automation system

A.S. Chajkovskaja, V.V. Baranyuk

Abstract – Meta-description is necessary to organize, share and operate with information resources in the automation system. But standard universal meta description systems are not well suitable for classifiers. The paper analyzes classifiers' specifics and offers appropriate meta-attributes set, based on Dublin Core standard. Meta-description includes some general attributes from this standard as well as such specific parameters, as classification and coding methods, developing, maintaining and affirming organizations, approval and implementation dates and so on. It provides the full and irredundant description for classifiers. Another point of interest for this research is a classifiers' verification. Mistakes in classifiers' content cause many problems for automation system personnel and software, for example, data equivocation and incorrectness, system failures and etc., and cause a system to provide fallacious solutions for its tasks. Such effects are impermissible in critical systems. Different verification aspects were examined including code length, structure and alphabet, changes by status field, code and UUID duplication and etc. All this were summarized to the set of stages for automated verification process. Its application increases classifiers' quality and lets automation system to produce correct authentic solutions.

Keywords – automation system, information resource, classifier, meta-description, classifier's maintenance, classifier's meta-description, classifier's verification.