ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО

Столяр Сергей Ефимович

КЛАССИФИКАЦИИ

Публикуемая статья является фрагментом книги «Информатика: представление данных и алгоритмы», готовящейся к печати в издательстве «Невский Диалект». Подробнее о книге см. http://rain.ifmo.ru/cat/view.php/books/seva-2007.

Правильная классификация – одна из первых ступеней научного описания. От правильности классификации зависит и правильность дальнейшего изучения. Но, хотя классификация и ложится в основу всякого изучения, сама она должна быть результатом известной предварительной проработки.

В.Я. Пропп [1]

Хранение больших массивов информации должно предусматривать удобные механизмы обработки поисковых запросов, формулируемые средствами информационно-поискового языка (ИПЯ). Учитывая назначение и функции ИПЯ, необходимо обеспечить надлежащее структурирование данных, поскольку любой ИПЯ опирается на соответствующую классификацию.

Классификация есть система деления понятий (классов) на основе соподчинения. *Класс* (или род), объединяющий некоторую совокупность элементов, разделяется на подклассы (или виды) согласно признаку, меняющему свое значение при смене подкласса. Если в качестве множества значений признака принимается лишь один вариант и его отрицание, то мы имеем дело с *дихотомической классификацией*.

Дихотомической является самая известная в истории наук классификация – «древо Порфирия»[2]. Нисходя по древу Порфирия, мы наблюдаем увеличение количества видовых отличий (рисунок 1). Понятие «человек» дихотомически уже не делится, а представляет совокупность, состоящую из индивидов, которые различаются по случайным признакам, индивидуальным отличиям. Например, одним из индивидуальных отличий у Сократа[3] является его лысина на голове.

Кстати ——————————————>
Ныне проблема перечисления индивидуальных отличий решается на государственном уровне – выдачей гражданину паспорта. Не так давно в нашей стране нашли более радикальное решение, введя ИНН, идентификационный номер налогоплательщика. Это

[1] *Владимир Яковлевич Пропп* (1895–1970) – русский фольклорист, один из основоположников современной теории текста. // Цит. по [5, с. 7].

[2] *Порфирий* (233(?)–304(?)) – греческий философ-неоплатоник. Проблема универсалий и родо-видового деления обсуждается в его трактате «Введение в “Категории” Аристотеля».

[3] *Сократ* (470–399 до н.э.), древнегреческий философ.

двенадцатизначный цифровой код вида NNNNXXXXXXCC, в котором начальные цифры NNNN соответствуют коду налогового органа, присвоившего этот номер, XXXXXX – собственно порядковый номер записи о физическом лице в территориальном разделе единого государственного реестра, CC – рассчитанное по специальному алгоритму двузначное контрольное число.

Хранение больших массивов информации должно предусматривать удобные механизмы обработки поисковых запросов...

Порфирий указывает, что при делении понятия в качестве признака можно выбрать как собственный (устойчивый) признак, так и случайный (неустойчивый). Теперь-то мы понимаем, что объективный выбор устойчивого признака не является простой задачей, и степень объективности зависит от уровня развития науки, а вовсе не от наличия растительности на голове. Вот несколько тому примеров.

Пример 1.

Шведский естествоиспытатель Карл Линней[1] создал классификацию Systema naturae, разделив мир природы на три царства: неорганическое, растительное и животное, – а далее использовав такие уровни: классы, отряды, роды и виды. При этом он впервые (1753) применил так называемую *бинарную номенклатуру* (предложенную, впрочем, задолго до того[2]). Согласно ей, каждый вид обозначается парой латинских слов: сначала – род (с большой буквы) в единственном числе, затем – видовое название.

Классификация Линнея не выдержала испытания временем. С позиций современной биологии, система родовидовых отношений живого мира выглядит гораздо сложнее: домен (надцарство) – царство – тип (отдел) – класс – отряд (порядок) – семейство – род – вид. Напомним читателю, что он относится к домену *эукариоты*, царству *животные*, типу *хордовые*, подтипу *позвоночные*, классу *млекопитающие*, подклассу *плацентарные*, отряду *приматы*, подотряду *высшие обезьяны*, семейству *человекообразные*, роду *люди (homo)* и, наконец, гордо именуемому виду *Homo sapiens*. Впрочем, нет полной уверенности, что нас куда-нибудь не переместят, поскольку, согласно авторитетному источнику[3], «разные исследователи выделяют от 4 до 26 различных царств, типов – от 33 до 132, классов – от 100 до 200». Кстати говоря, если бы на голове Сократа была шерсть, то владельца растительности по этому характерному признаку на-

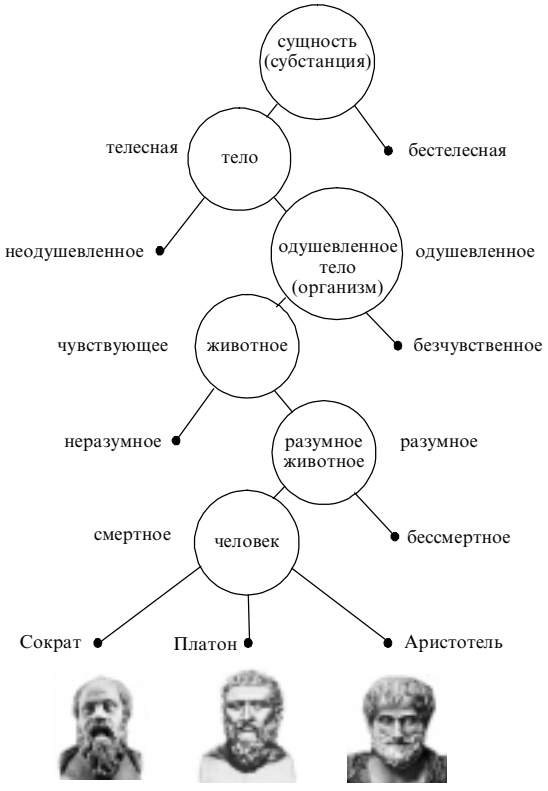
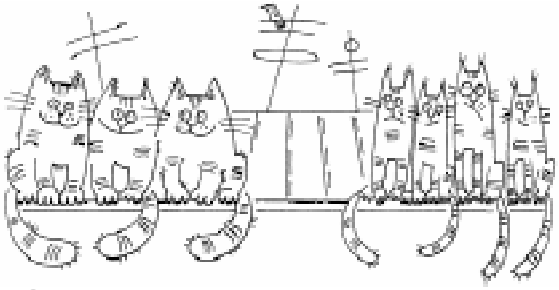


Рисунок 1. Древо Порфирия.

[1] *Carolus Linnaeus* (1707–1778), с 1762 г. Carl Linne, иностранный почетный член Петербургской АН (1754).

[2] *K. Baugin*, 1620.

[3] http://bio.fizteh.ru/student/files/biology/biolections



Необоснованный выбор оснований деления понятий (признаков) чреват дальнейшими ошибками.

верняка можно было бы отнести, по крайней мере, к классу млекопитающих, поскольку этот «классовый признак» современная биология, в отличие от представлений Порфирия, считает устойчивым.

Пример 2.

Широко и заслуженно признанной классификацией химических элементов является ныне *Периодический закон*, впервые представленный Д.И. Менделеевым[1] научному сообществу в 1869 г.

Однако к указанному времени уже существовало около 50 конкурирующих классификаций, большинство из которых также основывались на табличной форме. Любопытно замечание, обращенное к английскому ученому Дж.Ньюлендсу[2] одним из его оппонентов на заседании Лондонского химического общества (1865): «Не пробовал ли докладчик расположить в таблице элементы в алфавитном порядке и не заметил ли при таком расположении каких-либо новых закономерностей?»[3] Впрочем, встречались и оригинальные геометрические конструкции, вроде цилиндрической спирали («теллуровый винт») А. де Шанкуртуа[4] (1862).

Классификации, вводимые для описания объектов мира природы, принято именовать *естественными*. Необоснованный выбор оснований деления понятий (признаков) чреват дальнейшими ошибками. Так, следуя классификации Порфирия, к людям («разумным животным») приходится причислить и гуигнгнмов[5]. Или: в «октавной» классификации химических элементов Дж.Ньюлендса на несколько позиций претендовали по два элемента одновременно.

Альтернативу естественным представляют классификации *искусственные*, предназначенные обычно для систематизации создаваемых человеком предметов. Здесь исследователь гораздо менее ограничен в выборе признаков деления.

Пример 3.

Известно немало классификаций музыкальных инструментов.[6] В Китае издавна делили их на четыре группы, учитывая материал: каменные, деревянные, шелковые, металлические. В Древней Индии классификация иная: струнные инструменты, духовые инструменты, ударные инструменты из дерева или металла и, отдельно, ударные инструменты с кожаной мембраной (барабаны). Европейская классическая традиция выделяет три основных типа: духовые, струнные и ударные, к которым в наше время добавляются электронные инструменты.

Есть ли практическое значение, например, у последней из классификаций? Судите сами: в консерватории такая типизация находит отражение в делении на учебные кафедры, в оркестре по аналогичной схеме рассаживаются исполнители.

Очевидно, задачи классифицирования представляют практический интерес. Что же касается теоретического аспекта, то он проявляется как в исследованиях, посвящаемых классификации конкретных предметных областей, так и в фундаментальных работах на тему «классификации классификаций».

Наиболее простой классификацией, как мы уже видели выше, является дихотомическое дерево, в котором деление класса производится по некоторому признаку ровно на два подкласса. Соответствующая конструкция оператора ветвления **`if-then-else`** знако-

---

[1] *Дмитрий Иванович Менделеев* (1834–1907), химик, чл.-корр. (1876) Петербургской АН.

[2] *Newlands, J.A.R.* (1837–1898). Первым предложил термин «порядковый номер» элемента (1875).

[3] Цит. по [8, с. 22].

[4] *Chancourtois, A.E. Beguyer* (1819–1886), французский геохимик, профессор Парижской высшей горной школы.

[5] «...согласно английской орфографии его можно написать как houyhnhnm. Произношение этого слова давалось мне не так легко <...>, но после двух или трех попыток дело пошло лучше, и оба коня были, по-видимому, поражены моей смышленостью». – Цит. по [7, с. 147].

[6] http://metodolog.ru/00268/00268.html

ма любому программисту. Дихотомический механизм деления обобщается вполне естественным образом, если предположить, что класс может порождать более двух подклассов. Иначе говоря, признак может принимать несколько разных значений. Практическое значение такого разбиения также не обошли вниманием разработчики языков программирования, введя оператор выбора **case.**

Кстати ——————————————>

Одним из способов записи алгоритмов являются диаграммы Насси–Шнейдермана. Вид конструкции case на языке этих диаграм ясен из примера на рисунке 2.

<——————————————

Классификационные системы подобного рода относят к перечислительным [9]. Простейшие из них, «одноуровневые», называют *порядковыми.*

Пример 4.

Нумерация домов вдоль улицы – типичный пример перечислительной классификации. Порядок здесь, очевидно, линейный, определяемый последовательностью перечисления домов; хотя даже в этой простой типизации дополнительно выделены два подкласса – нечетные номера по правой стороне улицы, четные – по левой. При последующей застройке встраивание нового дома может нарушить имеющийся порядок, поэтому используют дополнительный индекс: так, между домами 3 и 5 появляется дом «3а» или «3, корпус 2».

Введение большего числа уровней делает перечислительную классификацию уже *иерархической*. Общее, делимое понятие соотносят с нулевым уровнем.

Пример 5.

Если в населенном пункте несколько улиц, то адрес дома формируется из названия улицы (первый уровень перечислительной классификации) и номера дома (второй уровень).

Пример 6.

Помещение, в котором размещается дисплейный класс, имеет порядковый номер 462. Это не значит, что в здании имеются помещения со всеми меньшими номерами, – в действительности начальная цифра 4 соответствует этажу, а номер 62 – порядковый для данного этажа. В данном случае можно предположить, что на этаже не более 99 помещений, либо привлекаются дополнительные индексы, как в примере 4.

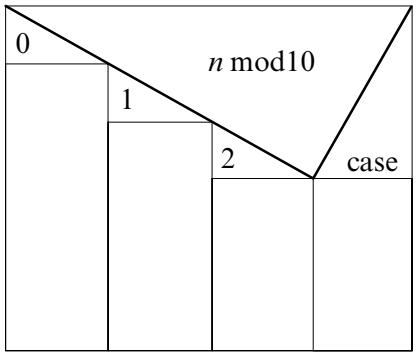


Рисунок 2. Блок **case**.

В отличие от дихотомического деления, для полного представления о множестве подклассов делимого класса следует либо указывать диапазон их значений, либо явно перечислять все значения. Наглядным способом представления иерархической классификации служит *иерархическое дерево*, естественным образом обобщающее вариант дерева дихотомического. Обычно перечисление вершин-детей требует определенного порядка. Чтобы локализовать место какого-то вида (значения) внутри иерархии, достаточно указать все родовые признаки деления в последовательности от прародителя к данному виду. Такая последовательность задает путь для реализации поискового запроса.

Пример 7.

В представленном на рисунке 3 дереве порядок «на детях» (подклассах) каждой вершины задается перечислением имен-значений тра-

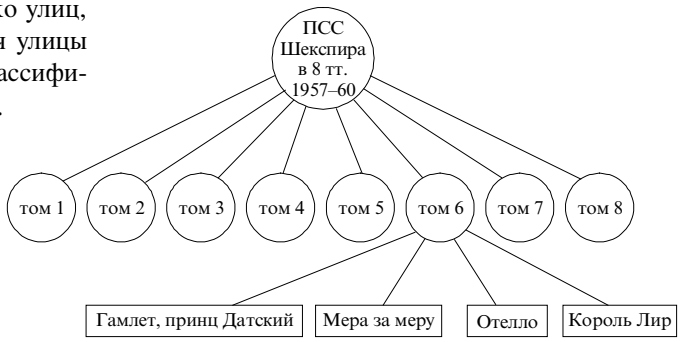


Рисунок 3. Иерархическое дерево.

...за классом А закреплена функция «вредительство», за классом Б — борьба с вредителем...

диционным способом «слева направо». Это иерархическое дерево иллюстрирует оглавление полного собрания сочинений Шекспира [10]. Место в нем пьесы «Гамлет, принц Датский» локализуется как

ПСС Шекспира || т. 6 || Гамлет, принц Датский

(сравните с соответствующей библиографической записью).

> ***Кстати*** ——————————————>
> Аналогичным образом формируется полное имя файла на диске.
> <——————————————

Далеко не всегда удобно использовать столь длинное имя. Зачастую проще оказывается ввести маркировку подклассов делимого класса элементами какого-либо общеупотребительного алфавита. В этом смысле любопытна классификация, представленная в примере 8.

Пример 8.

В.Я. Пропп выделил и классифицировал сюжетные элементы, из которых строится любая «волшебная сказка» [5]. Так, за классом **А** закреплена функция «вредительство», за классом **Б** – борьба с вредителем, **П** – победа над вредителем, **Л** – ликвидация беды или недостачи и т.д. Одного уровня не хватает, и каждый класс делится на подклассы, что отражается в наращивании индекса. В частности, для класса **А** это

А1 – похищение человека;
А2 – похищение волшебного средства или помощника;
АII – насильственное отнятие помощника;
А3 – порча посева;
А4 – похищение дневного света;
А5 – хищения в иных формах;
...
А16 – угроза насильственного супружества;
AXVI – то же между родственниками;
А17 – каннибализм или его угроза;
AXVII – то же между родственниками;
А18 – вампиризм (болезнь);
А19 – объявление войны.

Любопытно, что многим из перечисленных пунктов сей сказочной классификации соответствуют отдельные статьи вполне реального Уголовного кодекса. Собственно говоря, в основе законодательно принимаемого Уголовного кодекса лежит определенная классификация всевозможных преступлений.

Проппу явно не хватало набора десятичных номеров и он слегка «разбавил» перечисление, использовав заодно римскую нумерацию. Если в этом случае автор, специалист в области гуманитарных наук, несколько вольно обошелся с индексирующим алфавитом, то в более серьезных классификациях смешение алфавитов для индексирования элементов одного уровня не принято. Чаще других для индексирования привлекают либо латинский алфавит, либо десятичную нумерацию. Очевидно, популярность десятичного алфавита изначально объясняется исключительно антропологическими особенностями нашего организма. Алфавитный порядок значений индексов однозначно определяет последовательность подклассов-элементов внутри делимого класса.

Как мы успели заметить, для отдельных предметных областей обычно создаются частные классификации. Но для межпредметного применения, для документирования информации и организации ИПС они явно не подходят, здесь требуется уже универсальная классификация. Таковых в современной международной практике признано сразу несколько, и вполне обычным делом является их параллельное использование.[1]

В РФ действует ГОСТ 7.59-2003 [1], который допускает к применению, с уче-

[1] Пришлось специалистам в области классификаций совместно с компьютерщиками разработать специальные коммуникативные форматы UNIMARC, RUSMARC, MARC21, обеспечивающие адекватное конвертирование данных.

том вида и назначения документа, лишь семь классификационных ИПЯ, согласно определенному перечню. В числе прочих в него входят Универсальная десятичная классификация (УДК), Библиотечно-библиографическая классификация (ББК) и Десятичная классификация Дьюи (ДКД).

Читатель наверняка обращал внимание на специальную маркировку книг, размещаемую на обороте титульного листа, – в его левом верхнем углу, непосредственно перед библиографическим описанием, – один или два *классификационных индекса.* В отечественном книгоиздании принято в эту маркировку включать индексы УДК и ББК, они же дублируются в библиотечных каталожных карточках. Однако еще более распространенной является ДКД, ее индексами маркированы большинство зарубежных изданий.

Формально-логическое устройство ДКД выглядит достаточно просто. Дьюи принял за основу схему деления предметных областей, восходящую к Бэкону.

> ***Francis Bacon*** *(1561–1626), английский философ и политик. Согласно его классификации наук, «главными» являются история, поэзия, философия. Любопытно, что среди литературоведов уже лет триста то утихают, то вновь разгораются дискуссии относительно авторства шекспировских произведений, и среди альтернативных кандидатур фигурирует также сэр Фрэнсис Бэкон.*

Все пространство знания он поделил на 9 основных классов, занумерованных цифрами от 1 до 9 (поначалу 0 ему не понадобился, однако в дальнейшем нулевому классу нашли применение), основной класс – на 10 разделов, каждый раздел – на 10 секций.

Пример 9.

Нетрудно догадаться, что Дьюи не имел представления о компьютерных дисциплинах, так что свободный раздел **00** оказался весьма кстати: в нем нашли пристанище **Компьютеры, Интернет и системы**. Среди секций этого раздела есть, в частности, **004 = Обработка данных**. **Вычислительная техника**, а также **005 = Компьютерное программирование, программы, данные**.

> *В своих воспоминаниях американский библиотековед* ***Melville Louis Kossuth Dewey*** *(1851–1931) утверждал, что рождение ДКД стало результатом озарения, связанного с религиозным опытом. Так это или нет, уже не имеет значения, но со свойственной американцам практичностью Дьюи не преминул, публикуя «ниспосланные свыше» знания, пометить их собственным копирайтом. Разработка Dewey Decimal Classification датируется 1873 г., первое издание опубликовано в 1876 г. С тех пор ДКД постоянно обновляется, в 2003 г. вышло уже 22-е издание. В сферу распространения ДКД входят более 135 стран, доступны переводы на три с лишним десятка языков. В нашей стране опубликован перевод 21-го издания [2].*

Начальную триаду *класс–раздел–секция* при дальнейшем уточнении понятия принято отделять точкой, после которой справа добавляются новые цифры без формального ограничения их количества и без введения каких-либо синтаксических разделителей.

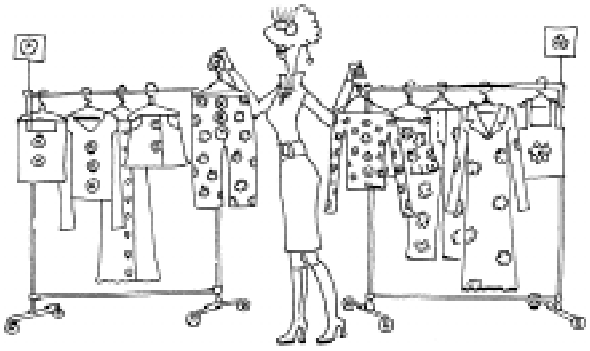
Пример 10.

В секции **746 = Роспись по ткани** имеется много-много подуровней, так **746.14043708997 = Индейские шерстяные одеяла**, а еще ниже **746.140437089972 = Шерстяные накидки индейцев навахо**.

Классификация не может оставаться статичной, вновь возникающие понятия не только требуют ее расширения, но и провоцируют противоречия в ее структуре. Потому очевидна необходимость регулярного обновления используемой версии ДКД, выпуска новых изданий. Существенно, что некоторые поправки вносятся и публикуются даже в период между переизданиями полной классификации.

Пример 11.

Секция **652 = Процессы письменного общения** понесла потери: понятие **652.5 = Word processing** получило в 22-м издании новую прописку – **005.52**. В нем же пополнилась секция **006 = Специальные компьютерные методы**: понятие **006.74 = Языки разметки** возникло совсем недавно.



...выбор признаков деления в искусственных классификациях достаточно произволен.

Даже из одного лишь примера 10 видно, как может разрастаться индекс, который приписывается некоторому понятию, нашедшему свое место в иерархической классификации. Чем больше признаков участвовало в последовательных делениях понятия, тем больше уровней учтено в полученном индексе. Выше мы отмечали, что выбор признаков деления в искусственных классификациях достаточно произволен. А значит, для некоторого индексированного понятия иная последовательность признаков деления породила бы другое значение индекса. Собственно говоря, пример 11 иллюстрирует подобную ситуацию. Таким образом, с ростом мощности множества вошедших в иерархическую классификацию понятий указанные недостатки все более тормозят ее расширение.

При иерархическом делении на каждом уровне учитывается лишь один признак, остальные рассматриваются как несущественные и отбрасываются. Почему бы не попытаться «параллельно» учесть некоторые из них? Новая идея состояла в том, чтобы собрать наборы одинаковых второстепенных признаков и строить на их основе дополнительные *таблицы типовых делений.* Соответствующие индексы должны добавляться к индексу основному. Разумеется, соединение нескольких значений в общем индексе понятия требует включения в ИПС расширенного синтаксиса. Подобные идеи нашли отражение в развитии ДКД, начиная с 13-го издания. С этого момента на смену иерархическим пришли более удобные *комбинационные классификации.*

В полной мере комбинационный подход проявился в Универсальной десятичной классификации (1895–1905), детище П.Отле и А.Лафонтена. В российской практике книгоиздания систематизация по УДК введена в 1963 г. Но еще задолго до официального признания и внедрения УДК, начиная с 30-х гг. прошлого века, в нашей стране разрабатывалась собственная система – ББК (1-е издание осуществлено в 1960–1968 гг.).

Дальнейшие шаги специалистов в области классификаций связаны с развитием идей Ш.Р. Ранганатана,[1] разработавшего основы фасетного анализа, специальный язык и фасетную систему классификации [6] (1-е изд. появилось в 1933 г.)[2] В современных версиях УДК и ББК «фасетность» тоже имеет место, но лишь «в малых дозах». Как мы отмечали, обе эти классификации применяются параллельно, в естественно-научных областях знания приоритет за первой из них, в остальных – за второй.

Пример 12.

Классификационные индексы книги [3] выглядят так: **УДК 681.142.2** и **ББК 32.973.26-018.2$_{Я}$75**; у книги [4] – **УДК 512+519.6** и **ББК 22.14+22.19**. Синтаксические связки–разделители «+», «–», «Я» сигнализируют о переключении таблицы типовых делений.

Как видим, все словарные операции актуальны для большинства из описанных

Бельгийские ученые-юристы ***Поль Отле*** *(Paul Otlet, 1868–1944) и* ***Анри Лафонтен*** *(Henri La Fontaine, 1854–1943) более известны как основатели (1895) Международного библиографического института (International Institute of Bibliography), после многих трансформаций превратившегося спустя столетие (1988) в Международную федерацию по [информации и] документации, МФД (International Federation for Information and Documentation, FID). Между прочим, каталожная карточка размером* 125 × 75 мм (5" × 3") *– тоже их изобретение.*

[1] *Shiyali Ramamrita Ranganathan* (1892–1972), индийский ученый.

[2] К сожалению, обсуждению этой интересной темы мы не можем уделить здесь достаточно места.

в этом разделе классификаций. Практичным вариантом алгоритмической структуры для их представления оказывается дерево, поскольку для него механизмы деления понятия и наследования признаков достаточно естественны.

Кстати ——————————————>

Древовидную структуру имеют иерархии классов в объектно-ориентированных языках программирования (при условии, что множественное наследование запрещено или не используется). В корне дерева находится самый общий класс (в Delphi – TObject), а классы-потомки наследуют его свойства и добавляют новые. В качестве примера приведем часть иерархии классов среды программирования Delphi рисунок 4.

<——————————————

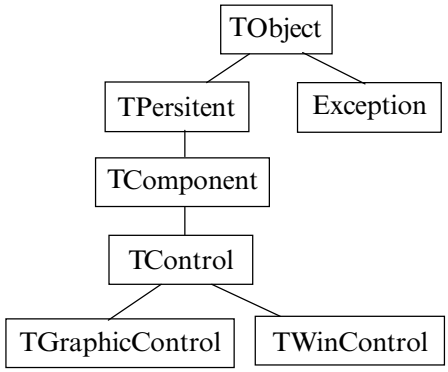


Рисунок 4. Иерархия классов среды программирования.

Литература

[1] ГОСТ 7.59-2003. Индексирование документов. Общие требования к систематизации и предметизации : Межгос. стандарт. Введ. 01.01.2004.

[2] Десятичная классификация Дьюи и Относительный указатель: В 4 т. / Пер. с англ. Под общ. рук. Я.Л.Шрайберга. Отв. ред. Е.М.Зайцева. 21-е изд. М.: ГПНТБ России, 2000.

[3] *Кнут Д.Э.* Искусство программирования. Т. 3. Сортировка и поиск. 2-е изд. М.: Вильямс, 2000.

[4] *Ноден П., Китте К.* Алгебраическая алгоритмика (с упражнениями и решениями). М.: Мир, 1999.

[5] *Пропп В.Я.* [Собр. трудов:] Морфология «волшебной» сказки. Исторические корни волшебной сказки. М.: Лабиринт, 1998.

[6] *Ранганатан Ш.Р.* Классификация двоеточием: Основная классификация. М.: ГПНТБ СССР, 1970.

[7] *Свифт Д.* Избранное. Л.: Худож. лит., 1987.

[8] *Семишин В.И.* Периодическая система элементов Д.И.Менделеева. М.: Химия, 1972.

[9] С*укиасян Э.Р.* Классификационные системы в их историческом развитии: проблемы терминологии и типологии // Науч. и техн. б-ки. 1998. № 11. С. 5–16.

[10] *Шекспир У.* Полн. собр. соч.: В 8т. Т. 6: Гамлет, принц Датский; Мера за меру; Отелло. М.: Искусство, 1960.

Столяр Сергей Ефимович,
учитель информатики,
лицей «Физико-техническая школа».