УДК 004.82:[002:004]

И. М. Зацман

Вербально-образное представление знаний в электронных библиотеках.* Ч. II

Рассматривается типология сфер представления знаний в электронной библиотеке, определяются такие понятия, как "семиотическая аппроксимация", "вербально-образный тезаурус", сравниваются конвенциональные основы построения вербальных и вербально-образных тезаурусов электронных библиотек.

1. ВВЕДЕНИЕ

В первой части статьи [1] предложена типология коммуникативных компонентов научных документов и показано, что графические компоненты обладают такими значениями семиотических характеристик, которые не свойственны вербальным и структурным компонентам. Рассмотренные значения семиотических характеристик компонентов научных документов были сгруппированы по следующим пяти позициям:

- детерминированность/размытость очертаний знаков компоненты,
- статика/динамика форм знаков,
- одномерность/двухмерность/многомерность форм знаков,
- упорядоченность/неупорядоченность сочетаний знаков,
- дискретность/континуальность сочетаний знаков.

В предлагаемой типологии семиотических характеристик первая характеристика для вербальных и невербальных компонентов может принимать следующие значения (возможные значения остальных характеристик следуют из их названий) [2]:

- однозначность выделения в компоненте знаков с детерминированными формами (означающими),
- многовариантность выделения знаков с детерминированными формами,
- присутствие в компоненте хотя бы одного нечеткого (размытого по очертаниям) знака.

В [1] показано, что для получения знаковых представлений научных документов электронной библиотеки, включая все их вербальные и невербальные компоненты, необходимо построить семиотическую систему библиотеки, включающую следующие системы знаков:

- традиционные вербальные системы знаков естественных языков (слова, устойчивые словосочетания и предложения),
- системы структурных знаков,
- системы графических знаков,
- системы неоднородных знаков (вербально-структурных, вербально-графических, структурно-графических и вербально-структурно-графических).

Этот перечень систем знаков соответствует типологии коммуникативных компонентов научных документов, включающей семь основных видов компонентов. В этот перечень, кроме систем знаков однородных компонентов, включены также системы неоднородных знаков.

Основная цель второй части статьи заключается в описании понятия "семиотическая аппроксимация", как основы построения систем графических и неоднородных знаков документов электронной библиотеки. При этом учитывается, во-первых, что семантическое пространство электронной библиотеки является семиотически неоднородным. Семиотическая неоднородность выражается в том, что ее области, соответствующие компонентам семи видов — вербальная, структурная, графическая, вербально-структурная, вербально-графическая, структурно-графическая и вербально-структурно-графическая — могут быть моноязычными, мультиязычными, а также областями с непрерывными языковыми системами и с неопределенной языковой принадлежностью компонентов.

Во-вторых, ключевой характеристикой каждой компоненты является ее *модальность*, под которой понимается принадлежность знаков компоненты некоторой языковой и/или вербально-образной классификационной системам: вербальная модальность знаков — естественные языки, математическая — язык математических формул, химическая — язык структурных химических формул, картографическая — вербально-образная классификация топографических и тематических карт, а также изображенных на них объектов и явлений и т. д. Компонент, включающий знаки разных языковых или классификационных систем, будем называть *полимодальным*. Компонент, включающий знаки с нечеткой модальностью, будем называть *неопределенным по модальности*. Знаковые системы электронных библиотек, включающих научные документы с компонентами разных модальностей, с полимодальными и неопределенными по модальности компонентами, будем называть *мультимодальными семиотическими системами*.

Исследование взаимосвязей языковой принадлежности и модальности, которые во многом зависят от вида компонента, не является целью этой статьи. Однако для дальнейшего рассуждения важно зафиксировать существование графических компонентов документов, о языке которых ничего не

* Работа выполнена при частичной поддержке РФФИ в рамках проекта № 01-06-80332.

известно, но при этом известна их тематика, отраженная в некоторой конвенциональной классификационной системе. Примером подобных графических компонентов могут служить изображения литолого-стратиграфических разрезов [3, с. 46–48].

Описание семиотической аппроксимации дается на примере тех графических компонентов, первая семиотическая характеристика которых может принимать только первые два значения: однозначность или многовариантность выделения в компоненте знаков с детерминированными формами, а вторая — может принимать только одно значение "статика", т. е. форма знаков во времени не меняется. Что касается остальных семиотических характеристик, то они могут принимать любые значения.

2. СЕМИОТИЧЕСКАЯ АППРОКСИМАЦИЯ: ПОСТАНОВКА ПРОБЛЕМЫ

Введение понятия "семиотическая аппроксимация" — это по сути еще одна попытка определить некоторую систему базовых элементов для построения графических образов. Результатом предыдущих попыток стало утверждение, что определить базовые элементы для построения графических образов подобно тому, как определяются буквы алфавита, иероглифы и вербальные знаки, составляющие вербальные тексты, невозможно [4].

В первой части статьи показано, что графические компоненты научных документов, которые являются континуумом точек, формально допускают бесконечное число вариантов их знакового представления в логико-семантических моделях документов [1, 5]. В этой ситуации, когда точное и однозначное определение базовых элементов для построения графических образов в семиотике считается невозможным, предлагается искать приближенное решение.

В данном случае приближенность решения понимается в трех аспектах. Во-первых, базовый элемент не должен обязательно точно совпадать с каким-то из фрагментов графического компонента. Во-вторых, допускается ситуация, когда в графическом компоненте могут существовать фрагменты, содержательные аспекты которых не отражены в базовых элементах с однозначными детерминированными формами или, если базовые элементы имеют многовариантные формы, то содержательные аспекты могут не найти своего отражения ни в одном из возможных вариантов. В-третьих, отношения между базовыми элементами могут не отражать всю полноту семантических отношений, которую можно наблюдать в графическом компоненте. Таким образом, предлагаемая трактовка "приближенности" ориентирована на решение задач семантического поиска невербальной информации в электронных библиотеках, а не на максимально точное воспроизведение графического компонента как сочетания базовых элементов.

В постановке проблемы семиотической аппроксимации используются следующие понятия: знак-множество, метазнак, знаковый базис и вербально-образный тезаурус. Введение понятия "знак-множество" является обобщением понятия "знак", принятого в классической семиотике [6], за счет расширения спектра значений первой семиотической характеристики компонентов научных документов электронной библиотеки. Рассматривая знаки-множества как множества с математической точки зрения, можно оперировать с детерминированными и размытыми множествами, статичными и динамичными, одномерными и многомерными, дискретными, континуальными и дискретно-континуальными, упорядоченными и неупорядоченными, используя при этом существующий аппарат теории множеств. С семиотической точки зрения, рассматривая знаки-множества как знаки, можно использовать значения этих знаков (означаемое) для вербально-образного представления знаний в электронных библиотеках.

Введение понятия "знак-множество" дает возможность строить мультимодальные семиотические системы, включающие электронные знаки, формы которых (означающее) могут меняться в более широком спектре, чем традиционные знаки бумажных документов. Однако использование знаков-множеств человеком иногда может быть затруднено. Например, динамичные, трехмерные и размытые по очертаниям знаки-множества могут быть удобны для адекватного "фигуративного" представления в цифровой форме непрерывного изменения за некоторый период времени облачности в электронной климатической карте. Но даже приблизительное описание этих изменяющихся во времени знаков-множеств в традиционной научной статье на бумаге потребует как минимум дискретно-временного ряда картинок, изображающих в проекции состояние в отдельные моменты времени этих объемных знаков с размытыми очертаниями. Поэтому в традиционных документах предлагается ввести понятие "метазнак" как условное обозначение знака-множества в виде индексированных букв алфавита. Возможно использовать метазнаки и при компьютерной обработке информации в качестве указателей на места хранения знаков-множеств. С помощью введения метазнаков в языки семантической разметки можно распространить сферу их применения на графические компоненты документов. В этом случае в линейные конструкции языков семантической разметки, используемые человеком, предлагается включать метазнаки как указатели на электронные знаки-множества. А при обработке цифровых кодов этих конструкций компьютеру становятся доступны и сами знаки-множества.

Взаимосвязь между метазнаком, знаком-множеством, денотатом и концептом схематично будем изображать в виде тетраэдра (рис. 1), являющегося естественным обобщением семиотического треугольника Фреге.

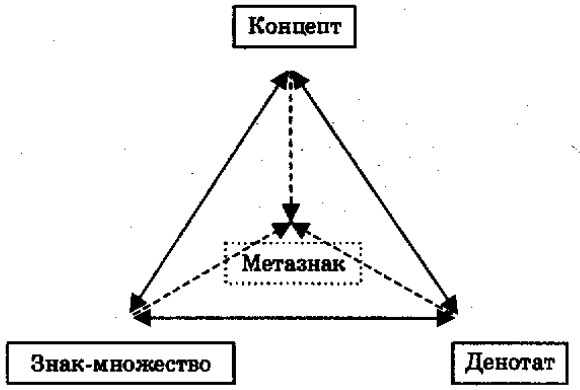


Рис. 1. Семиотический тетраэдр

Рассматривая в качестве базовых элементов знаки-множества, определим *семиотическую аппроксимацию* графического компонента научного

документа как семантическое соответствие, установленное между его континуумом точек и конечным числом заранее построенных знаков-множеств и их сочетаний.

В этом эмпирическом определении остаются открытыми следующие принципиальные вопросы: каким образом заранее строятся знаки-множества, используемые для семиотической аппроксимации, как устанавливается семантическое соответствие и измеряется мера этого соответствия. Эти вопросы предлагается рассматривать как постановку проблемы семиотической аппроксимации графических компонентов с помощью знаков-множеств.

В качестве частного случая рассмотрим тематически однородные графические компоненты документов некоторой электронной библиотеки. Например, если речь идет о картах, то предполагается, что известна их тематическая классификация (геологические карты, почвенные, климатические и т. д.), в соответствие с которой карты могут быть упорядочены. Предположим, что для тематически однородных компонентов можно указать перечень классов объектов, которые эти компоненты включают. Например, для топографической карты можно указать основные классы составляющих ее объектов (формы рельефа, элементы гидрографической сети и т. д.).

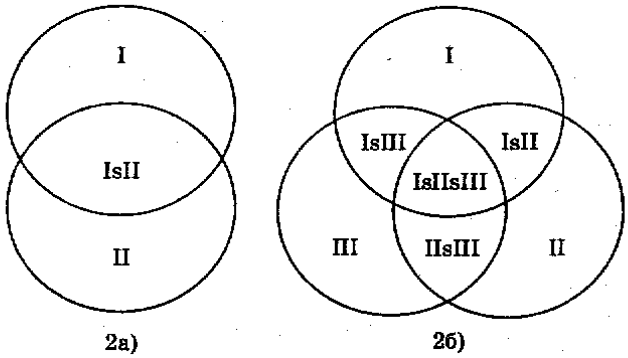
Тогда сочетание знаков-множеств, построенное для любого класса объектов и явлений, отраженных в тематически однородных графических компонентах документов электронной библиотеки на основе выбора знаков-множеств из имеющихся дескрипторов вербально-образного тезауруса этой библиотеки, будем называть *знаковым базисом* класса в том случае, если выбранные знаки-множества отражают содержательные аспекты объектов этого класса.

Приведенное определение знакового базиса для частного случая тематически однородных графических компонентов документов некоторой электронной библиотеки говорит о том, что одним из возможных источников формирования знаковых базисов классов может быть *вербально-образный тезаурус библиотеки*, который интегрирует семантическое описание вербальной и невербальной информации электронной библиотеки, структурированное на основе некоторой типологии сфер представления знаний. Для этого частного случая включим в постановку проблемы семиотической аппроксимации графических компонентов вопрос о конвенциональной основе и принципах построения вербально-образного тезауруса электронной библиотеки.

3. ТИПОЛОГИЯ СФЕР ПРЕДСТАВЛЕНИЯ ЗНАНИЙ В ЭЛЕКТРОННОЙ БИБЛИОТЕКЕ

Предлагаемая типология основана на следующем утверждении о существовании трех основных сфер представления знаний: вербальные знания в лингвистической форме, которые не могут быть адекватно переведены в невербальную форму; невербальные (нелингвистические) знания, которые не могут быть представлены в вербальной форме; и та часть знаний, которая может быть достаточно адекватно представлена и в вербальной, и в невербальной формах [6].

В работе [6] отношения между тремя сферами представления знаний иллюстрируются в виде двух частично пересекающихся окружностей (рис. 2а). Первая окружность условно обозначает сферу представления знаний в лингвистической форме (сфера I), а вторая — сферу представления знаний в нелингвистической форме (сфера II). Область, образованная пересечением этих окружностей, соответствует третьей сфере знаний, которые могут быть достаточно адекватно представлены в двух вариантах: и в лингвистической, и в нелингвистической формах.



2а) 2б)

Рис. 2. Сферы представления знаний в лингвистической и нелингвистической формах (а); в вербальной, структурной и графической формах [обозначены 7 из 11 возможных сфер] (б)

Отметим, что в теории множеств традиционно на примере области пересечения двух окружностей иллюстрируют операцию пересечения двух множеств, т. е. те элементы, которые принадлежат одновременно двум множествам. Однако на рис. 2а пересечение двух окружностей обозначает элементы третьего множества, которые не принадлежат первым двум. Поэтому рис. 2 надо рассматривать только с семиотической точки зрения в соответствии с типологией сфер знаний из работы [6], а не с теоретико-множественной точки зрения.

Для обозначения третьей сферы, которую будем называть сферой семиотической синонимии, предлагается использовать латинские цифры I, II и латинскую букву s между ними, которая является первой буквой в слове synonymy.

В статье из всех возможных нелингвистических форм рассматриваются структурные, графические формы представления знаний в документах и сочетание этих форм, в том числе с вербальной формой представления. Поэтому для знаний, отображаемых в электронных библиотеках научных документов, предлагается использовать более детальную типологию, которую назовем *таксономией или систематикой сфер вербально-образного представления знаний.*

Рассмотрим семь сфер, которые схематично обозначены на рис. 2б в виде семи областей, образованных тремя частично пересекающимися окружностями. Эти три окружности соответствуют вербальной, структурной и графической сферам вербально-образного представления знаний.

Проведем соответствие между таксономией сфер и типологией компонентов научных документов. Трем видам однородных компонентов документов (вербальным, структурным и графическим) соответствуют три сферы представления знаний. Отметим, что, выделяя только вербальную и невербальную сферы, получаем одну сферу семиотической синонимии. А при использовании таксономии, т. е. выделении вербальной (I), структурной

(II) и графической (III) сфер представления знаний, получаем четыре сферы семиотической синонимии (обозначены на рис. 2б, как IsII, IsIII, IIsIII и IsIIsIII).

Например, сегмент IsIIsIII обозначает ту сферу семиотической синонимии, знания в которой могут быть достаточно адекватно отражены в каждой из форм: вербальной, структурной и графической. В отличие от трех разных вариантов представления одних и тех же знаний (сегмент IsIIsIII) содержательные аспекты вербально-структурно-графических компонентов документов представляются через **одновременное** сочетание или вложенность трех однородных видов компонентов [1]. Отсюда следует, что те сферы знаний, которые представлены в научных документах неоднородными компонентами, не нашли свое отражение на рис. 2б.

Таким образом, на рис. 2б обозначены семь сфер вербально-образного представления знаний, но при этом не показаны еще четыре сферы, которые соответствуют четырем видам неоднородных компонентов документов электронной библиотеки, а именно: вербально-структурные, вербально-графические, структурно-графические и вербально-структурно-графические, определенные в первой части статьи [1].

Для того, чтобы показать и эти четыре сферы представления знаний, отойдем от иллюстративной методики из работы [6], в соответствии с которой в центре рис. 2а расположена область семиотической синонимии IsII, и изобразим все сферы семиотической синонимии с внешней стороны окружностей. На рис. 3 даны 11 сфер представления знаний с выделением для сфер IsII, IsIII, IIsIII и IsIIsIII внешних к трем окружностям областей.

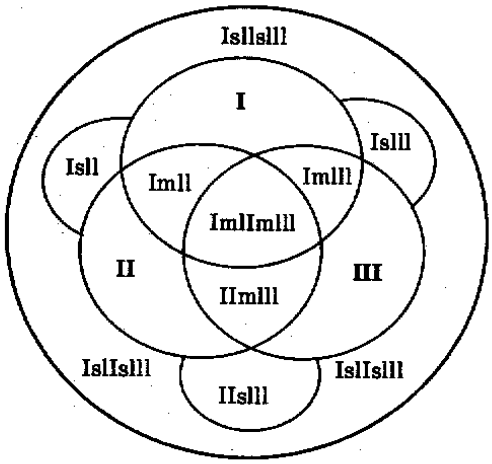


Рис. 3. Таксономия сфер вербально-образного представления знаний в электронных библиотеках научных документов

Это дает возможность изобразить еще четыре области ImII, ImIII, IImIII и ImIImIII, которые схематично представляют четыре сферы представления знаний, которые по аналогии с неоднородными компонентами будем называть: вербально-структурная, вербально-графическая, структурно-графическая и вербально-структурно-графическая сферы вербально-образного представления знаний. В обозначениях этих областей используется первая буква слова miscellaneous — смешанный или разнообразный.

Таксономия из одиннадцати сфер в наиболее общей форме отражает структуру семантического пространства электронной библиотеки научных документов. Частные случаи схематичного представления семантического пространства, включающего только однородные компоненты научных документов, рассмотрены в первой части статьи (см. рис. 4 и рис. 5 в работе [1]).

4. ВЕРБАЛЬНО-ОБРАЗНЫЙ ТЕЗАУРУС ЭЛЕКТРОННОЙ БИБЛИОТЕКИ

Рассмотренную таксономию сфер представления знаний предполагается использовать в дальнейших исследованиях как основу для решения проблемы семиотической аппроксимации и описания принципов построения вербально-образного тезауруса. В соответствие с этой таксономией дескрипторы вербально-образного тезауруса можно разделить, соответственно, на семь видов. Т. е. семи сферам вербально-образного представления знаний, обозначенным на рис. 3, как I, II, III, ImII, ImIII, IImIII и ImIImIII, поставим в соответствие семь видов дескрипторов: вербальные, структурные, графические, вербально-структурные, вербально-графические, структурно-графические и вербально-структурно-графические. Например, графическая часть тезауруса включает графические дескрипторы и именно из нее предлагается выбирать графические знаки-множества для построения знаковых базисов графических компонентов научных документов электронной библиотеки.

Для обозначения вербально-структурных, вербально-графических и вербально-структурно-графических дескрипторов будем использовать термин *вербально-образные дескрипторы*, а для структурных, графических и структурно-графических дескрипторов — *образные дескрипторы*.

В традиционных вербальных тезаурусах имеется три основных типа связей и отношений, которые естественно сохранить и в вербально-образном тезаурусе: (а) предпочтительные, на основе которых выделяется дескриптор из соответствующего синонимического ряда знаков; (б) иерархические; (в) ассоциативные: целое — часть, часть — целое, причина — следствие, следствие — причина, функциональное сходство. В вербальных тезаурусах ряды синонимов, в том числе многословных, могут иметь существительные и именные группы, прилагательные, глаголы и глагольные группы [7, 8].

Включение в тезаурус образных и вербально-образных дескрипторов влечет существенное расширение типов отношений между дескрипторами. В первую очередь, в вербально-образном тезаурусе необходим принципиально новый тип отношений семиотической синонимии. На рис. 2б и рис. 3 обозначены четыре сферы IsII, IsIII, IIsIII и IsIIsIII с этим типом отношений, которым соответствуют четыре логических области тезауруса. Эти области могут включать дескрипторы разных видов (вербальные, образные или вербально-образные) при наличии между ними отношений семиотической синонимии. Семиотические синонимы-дескрипторы отличаются от традиционных вербальных синонимов-дескрипторов тем, что традиционные синонимы относятся к одной вербальной сфере представления знаний и имеют одинаковую вербальную модальность.

Рассмотрим явление семиотической синонимии на примере химической номенклатуры, которая представляет собой совокупность названий индивидуальных химических веществ, их групп и классов, а также правила составления этих названий. При составлении названия на основе структурной химической формулы, как и при переводе названия в структурную формулу, последовательно выполняется набор формальных правил, а затем выявляется и называется родовая структура, к которой примыкает основная характеристическая группа соединения. При составлении названия используется классификация соединений и названия характеристических групп органических соединений [9].

В некоторых программах-редакторах структурной химической информации имеется возможность автоматического составления названий химических соединений на основе структурной химической формулы. На рис. 4 приведен пример структурной химической формулы вещества, название которого составлено редактором ISIS Draw 2.3 с дополнительной программой AutoNom Standard.

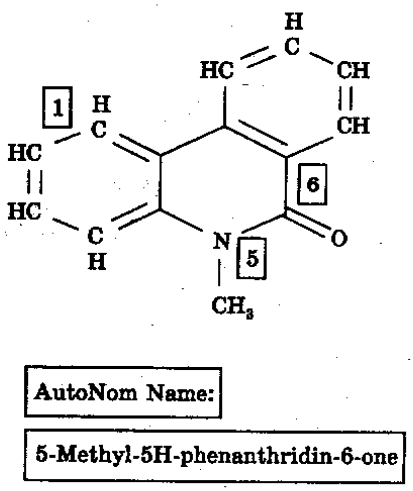


Рис. 4. Структурная формула и название одного и того же химического вещества

В приведенном примере в качестве семиотических синонимов выступают структурная формула и название химического вещества. Если они включены в вербально-образный тезаурус, соответственно, как структурный и вербальный дескрипторы-синонимы, то логически они принадлежат той области тезауруса, которая обозначена, как IsII. Отметим, что в научных документах подобные концепты могут быть представлены вербальными и/или структурными знаками и знаковыми образованиями.

Отношения семиотической синонимии в вербально-образном тезаурусе являются основой для решения задач обобщенного семантического поиска в электронных библиотеках, в которых интересующие пользователя сведения об искомых объектах, веществах или явлениях могут быть представлены в компонентах разных видов и ему не известна заранее компонентная форма интересующей его информации. Если в электронной библиотеке есть вербально-образный тезаурус с отношениями семиотической синонимии, то в запросе можно использовать любую из возможных компонентных форм представления для поиска интересующих его сведений **независимо от формы их представления** в искомых научных документах.

Кроме отношений семиотической синонимии в вербально-образном тезаурусе предлагается учитывать параметрические отношения между дескрипторами. На стадии декомпозиции научного документа на компоненты ранее рассматривалась возможность использования параметрической схемы структуризации [1, 10], которую можно рассматривать в качестве аналога для параметрических связей между дескрипторами.

Схемы параметрической структуризации применяются для декомпозиции документов на компоненты, которые в логической структуре документа связаны через значения одного или нескольких параметров. Параметрические связи между дескрипторами используются при поиске документов. В качестве примера параметрически связанных графических дескрипторов назовем серию знаков-множеств, соответствующих одному и тому же изменяющемуся геообъекту в разные моменты времени, т. е. параметром в данном случае является геологическое время. Тогда наличие параметрических отношений в вербально-образном тезаурусе для этого геообъекта является основой для организации поиска документов, содержащих сведения об этом геообъекте, с учетом информации в тезаурусе об изменении геообъекта во времени.

Отметим также, что в науках о Земле графический образ геообъекта в карте зависит от ее масштаба. Существуют правила генерализации карт, которые и определяют способы изменения форм знаков и знаковых образований на карте при изменении масштаба. Для описания отношений между знаками, которые соответствуют одному и тому же геообъекту, но используются на картах разных масштабов, предлагается ввести в вербально-образном тезаурусе отношения генерализации.

Отметим еще одно отличие вербально-образных тезаурусов от традиционных. Концепция построения традиционных тезаурусов основана на грамматическом самоописании естественных языков и достаточно четких границах между ними, а в концепции построения вербально-образных тезаурусов необходимо отразить реально существующую размытость границ между языками в системах невербальных знаков или отсутствие языков. При этом базовые элементы естественных языков (буквы, иероглифы, а также слова, устойчивые словосочетания и другие вербальные знаки) существуют независимо от тезаурусов электронных библиотек. Существенная часть базовых элементов графических компонентов (знаки-множества и устойчивые знаковые образования) определяются на основе дескрипторов вербально-образного тезауруса и зависят от степени отражения в этом тезаурусе конвенциональных правил построения графических компонентов научных документов. Простейшие базовые элементы в виде геометрических примитивов (точки, отрезки, окружности и т. д.), встречающиеся в графических компонентах, могут являться составными элементами знаков-множеств или нести некоторую семантическую нагрузку в сочетаниях со знаками-множествами.

5. КОНВЕНЦИОНАЛЬНАЯ ОСНОВА ПОСТРОЕНИЯ ТЕЗАУРУСОВ

Обозначив отдельные типы связей и отношений в вербально-образном тезаурусе, вернемся к вопросам нормализации, многозначности и недетерминированности выделения знаков в графических компонентах, определения языковой принадлежности

графических компонентов и составляющих их знаков.

Знаки-множества, входящие в базис любого класса объектов тематически однородных графических компонентов, по определению, являются графическими знаками, построенными на основе дескрипторов вербально-образного тезауруса конкретной электронной библиотеки. Таким образом, вербальные знаки соотнесены с некоторым вербальным языком, а графические знаки — с некоторым тезаурусом.

В случае вербального языка известны грамматическое описание языка, системы парадигматических, синтагматических и семантических отношений, конвенциональный характер которых учитывается при проектировании вербальных тезаурусов и разработке методов вербального поиска документов в электронных библиотеках.

Естественно попытаться очертить те научные знания, конвенциональность которых можно было бы использовать при проектировании вербально-образных тезаурусов и разработке методов семантического поиска в электронных библиотеках научных документов. При этом, как отмечалось выше, базовых элементов в визуальных (графических) языках не существует и, следовательно, конвенциональность базовых элементов, аналогичных буквам алфавитов и знакам естественных языков, отсутствует. Это говорит о необходимости поиска "неязыковой" конвенциональной основы построения вербально-образных тезаурусов.

Вернемся к определению понятия знакового базиса, с точки зрения поиска основы построения тезауруса. Это определение содержит следующие слова: "для любого класса объектов и явлений, отраженных в тематически однородных графических компонентах документов, хранящихся в электронной библиотеке". В этом определении подразумевается, во-первых, конвенциональная классификация объектов и явлений, знания о которых содержатся в научных документах. Во-вторых, это определение предполагает конвенциональную тематическую классификацию графических компонентов научных документов, отражающих научные знания в невербальной форме.

Поэтому в качестве конвенциональной основы построения вербально-образных тезаурусов предлагается использовать:

- существующие в научных областях знаний и научных специальностях вербально-образные системы классификаций объектов и явлений;
- используемые авторами научных документов традиционные классификации графических компонентов.

Например, в науках о Земле существует традиционное деление карт на топографические и тематические (геологические, геофизические, геохимические, климатические, экологические и т. д.).

Общность конвенциальных основ вербальных и вербально-образных тезаурусов заключается в использовании системы семантических отношений, которая для вербальных тезаурусов является конвенциональной системой семантических отношений естественного языка, а для вербально-образного тезауруса электронной библиотеки научных документов — системой семантических отношений соответствующих конвенциональных научных классификаций.

Конвенциональность основы построения вербально-образного тезауруса может быть использована для решения проблемы определения в документах электронной библиотеки значений авторских и слабоопределенных знаков (эти понятия введены в разделе 2 первой части статьи [1]). Строго говоря, только в случае определенных знаков корректно использовать термин *знак*, понимая определенный знак как относительно устойчивое в течение некоторого периода времени и общепринятое единство его формы (означающего) и значения (означаемого). Для авторских и слабоопределенных знаков единство формы и значения существует в пределах одного или нескольких документов, но не для всего корпуса документов электронной библиотеки.

В подобных случаях термин *знак* является не вполне корректным до тех пор, пока не определена (доопределена) относительно устойчивая система отношений и/или моделей, в рамках которой по форме знака может быть установлено его значение. Для эксплицитного установления связи формы и значения авторских и слабоопределенных знаков в семантическом пространстве электронной библиотеки можно использовать систему связей и отношений между дескрипторами ее вербально-образного тезауруса, которую предлагается строить на основе конвенциональных научных классификаций.

Чтобы доопределить авторские знаки некоторого документа необходимо устанавливать соответствие между этими знаками и теми дескрипторами вербально-образного тезауруса, которые соответствуют авторским знакам по смыслу документа. Несколько иной подход предлагается для слабоопределенных знаков. Если значения авторских знаков, как правило, определяется в документе, то смысловое содержание слабоопределенных знаков в научных документах иногда не раскрывается. В этом случае при формировании семантического пространства электронной библиотеки корпус научных документов надо дополнить материалами с описаниями используемых слабоопределенных знаков. Тогда для установления смыслового соответствия между слабоопределенными знаками и дескрипторами вербально-образного тезауруса может быть использовано содержание этих материалов.

Возможен случай, когда в тезаурусе не будет дескрипторов, необходимых для построения знаковых базисов, и следовательно, для доопределения авторских и слабоопределенных знаков. Эта ситуация известна и при ведении традиционных вербальных тезаурусов, когда необходимо их расширение.

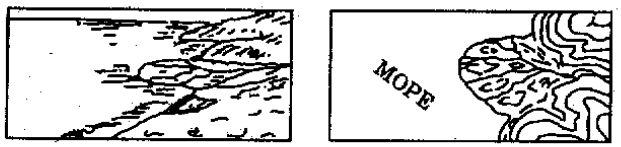
Если для дескрипторов вербально-образного тезауруса указывать соответствующие им классификационные системы и их классификационные идентификаторы, то это можно рассматривать как аналог указателя языковой принадлежности дескриптора вербального тезауруса. Таким образом, с помощью системы классификационных идентификаторов дескрипторов вербально-образного тезауруса определяется их классификационная принадлежность. Отсюда следует и возможность выявления принадлежности знаковых базисов классов объектов и явлений, отраженных в тематически однородных графических компонентах.

Существующая в некоторых классификационных системах размытость границ между отдельными категориями объектов и явлений в явной форме

может быть выражена через множественную отнесенность соответствующих дескрипторов к нескольким категориям.

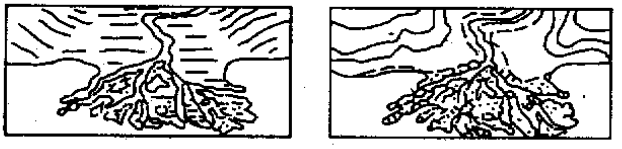
Предлагаемое обобщение понятия "знак" за счет введения понятия "знак-множество" в дальнейшем планируется использовать для построения знаковых базисов тематически неоднородных графических компонентов, а также для знакового представления неоднородных компонентов научных документов (вербально-структурных, вербально-графических, структурно-графических и вербально-структурно-графических).

В качестве иллюстрации направлений дальнейшего развития концепции вербально-образного представления знаний в электронных библиотеках научных документов применительно к вербально-графическим компонентам рассмотрим изображения нескольких форм рельефа (рис. 5 из [11]).



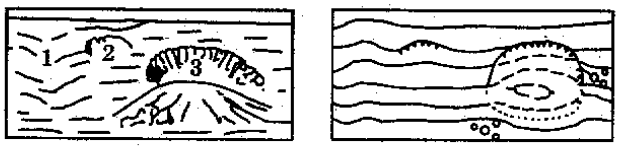
Дельта выступания

Дельта заполнения

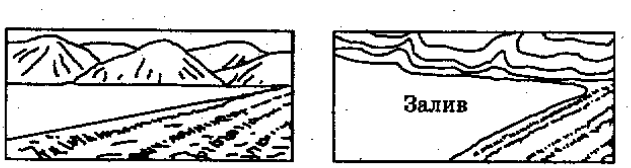


Дельта разветвления

Карстовые воронки



Суффозионная ложбина – 1; суффозионный цирк –2; оползень – 3



Надводная терраса с валами

Рис. 5. Примеры изображений форм рельефа [11]

Отметим, что содержательные аспекты отдельных форм рельефа передаются сочетанием слов и графики (берег моря и берег озера в этих примерах отличаются только по вербальным компонентам изображений). Для решения проблемы семантического поиска документов с изображениями форм рельефа на картографических компонентах необходимо иметь вербально-образный тезаурус, включающий достаточно полный перечень дескрипторов для основных форм рельефа и их составных частей.

Только после включения в тезаурус соответствующих дескрипторов появляется возможность для построения знаковых базисов для соответствующих классов объектов картографических компонентов. После построения знаковых базисов классов как систем графических и вербально-графических знаков появляется возможность решения задач знакового представления компонентов документов, включающих изображения форм рельефа.

В настоящее время классификационные системы в геоморфологии дают достаточно полную характеристику форм рельефа, включая семантические отношения между ними. Однако описание форм рельефа на знаковой основе, а тем более в виде сочетаний дескрипторов вербально-образного тезауруса не является задачей геоморфологии. Но геоморфология дает необходимую исходную информацию для проектирования соответствующих категорий дескрипторов.

Этот пример с формами рельефа иллюстрирует потенциал вербально-образных классификационных систем объектов и явлений, разработанных в различных областях знаний, который может быть использован при проектировании вербально-образного тезауруса и его наполнения вербальными, вербально-образными и образными дескрипторами. Вербальные дескрипторы необходимы для отражения в тезаурусе содержания верхних, достаточно общих и абстрактных уровней классификационных систем, где располагаются такие абстрактные термины как *элементы гидрографической сети, формы рельефа*. Для отражения в тезаурусе сведений о промежуточных и нижних уровнях вербально-образных классификационных систем объектов и явлений, которые сочетают текст и графику, могут потребоваться вербально-образные дескрипторы. К подобным изображениям на рис. 5 [11] относятся дельта выступания, включающая слово *море*, и дельта заполнения — слово *озеро*. Для отражения содержания уровней классификации, информация которых является чисто изобразительной, естественно использовать графические дескрипторы. На рис. 5 [11] дельта разветвления не имеет вербальных компонентов и для ее семиотической аппроксимации потребуются только графические дескрипторы.

Таким образом, тезаурус, интегрирующий вербально-образные классификационные системы различных научных специальностей, может быть основой для вербально-образного представления знаний в политематических электронных библиотеках научных документов.

6. ЗАКЛЮЧЕНИЕ

В двух частях статьи рассмотрены следующие составляющие концепции вербально-образного представления знаний:

- типология компонентов научных документов,
- описание семиотических характеристик и спектра принимаемых ими значений для вербальных и невербальных компонентов,
- таксономия сфер представления знаний в электронной библиотеке,
- семиотическая аппроксимация,
- конвенциональная основа построения вербально-образного тезауруса.

Показано, что семиотическая аппроксимация является основой построения мультимодальной семиотической системы электронной библиотеки,

включающей традиционные системы знаков электронных библиотек и знаковые базисы классов объектов и явлений, представленных в графических и неоднородных компонентах научных документов.

Знаковое представление вербально-образной информации с использованием мультимодальной семиотической системы, кроме ориентации на решение проблемы семантического поиска в электронных библиотеках, имеет самостоятельное значение, так как является основой для разработки языков обобщенной семантической разметки документов [12]. Известные в настоящее время языки и их приложения исходно не предназначены для разметки и кодирования континуальной и дискретно-континуальной графической информации.

Что касается широко используемого в настоящее время метаязыка XML, то можно говорить о практически полном отсутствии приложений XML, которые можно было бы использовать для разметки графической информации. Исключение составляет язык GML — Geography Markup Language — для разметки географической информации [13]. Это приложение метаязыка XML дает возможность конструировать простые картографические компоненты в виде наборов дискретно отделимых графических примитивов: линия, сочетание линий, полигон, сочетание полигонов. При этом каждый графический примитив определяется некоторым набором точек. Но этот язык не рассчитан на семантическую разметку и кодирование сложных континуальных картографических изображений.

Таким образом, в настоящее время, метаязык XML, его приложения не имеют синтаксических средств и конструкций для разметки и кодирования сложных графических изображений на знаковой основе. Можно предложить следующее направление развития языков разметки документов с использованием мультимодальной семиотической системы и вербально-образного тезауруса. При разметке документов со сложными графическими компонентами, в дополнение к ссылкам на их растровые изображения, конструкции языка обобщенной семантической разметки должны включать в документы метазнаки как ссылки на дескрипторы вербально-образного тезауруса. При компьютерной обработке документов метазнаки используются как указатели на места хранения в тезаурусе соответствующих дескрипторов. Введение метазнаков и новых конструкций в языки обобщенной семантической разметки дает возможность распространить сферу применения языков разметки на сложные графические компоненты.

Дескрипторы, как правило, не идентичны фрагментам графических компонентов, представленных в знаковой форме. Поэтому соответствующие им метазнаки в языках разметки используются как указатели на дескрипторы для решения задач поиска, а для визуализации и точного воспроизведения найденных документов могут использоваться растровые поточечные или векторные электронные формы графических компонентов.

В заключение отметим, что для понятий "знак-множество", "семиотическая аппроксимация" и "знаковый базис" в статье не ставилась задача их формального определения. Эти понятия используются как эвристические. При таком определении основных понятий отсутствует возможность теоретической оценки полноты и точности семиотической аппроксимации и семантического поиска. Однако остается возможность их экспертной оценки в каждом конкретном случае построения вариантов знаковых базисов и мультимодальной семиотической системы электронной библиотеки научных документов.

* * *

Автор выражает искреннюю признательность Ю. И. Шемакину, А. М. Курчавову и Т. Н. Херасковой за их предложения и замечания, сделанные в ходе подготовки статьи.

СПИСОК ЛИТЕРАТУРЫ

1. Зацман И. М. Вербально-образное представление знаний в электронных библиотеках. Ч. I // НТИ. Сер. 2. — 2001.— No 10 .— С. 20–30.

2. Зацман И. М. Семантическое кодирование и разметка геолого-географических документов в политематических электронных библиотеках // Информационные технологии.— 2000.— N 11.— С. 2–11.

3. Зайцев Ю. А., Хераскова Т. Н. Венд Центрального Казахстана.— М.: МГУ, 1979.

4. Jorna R. J., Heusden B. Signs, search and communication: Towords an empirical future for semiotics // Signs, search and communication: Semiotics aspects of artificial intelligence / Ed. R. J. Jorna, B. Heusden, R. Posner.— Berlin: Walter de Gruyter, 1993.— P. 1–21.

5. Зацман И. М. Семиотические проблемы моделирования и поиска полнотекстовых научных документов // Тр. Междунар. семинара Диалог-2001 по компьютерной лингвистике и ее приложениям, 30 мая–5 июня 2001 г. Т. 2.— Аксаково, 2001.— С. 136–144.

6. Eco U. A Theory of Semiotics.— Bloomington: Indiana University Press, 1976.— 356 pp.

7. Шемакин Ю. И. Тезаурус в автоматизированных системах управления и обработки информации.— М.: Воениздат, 1974.— 192 с.

8. Лукашевич Н. В., Добров Б. В. Тезаурус для автоматического концептуального индексирования как особый вид лингвистического ресурса // Тр. Междунар. семинара Диалог-2001 по компьютерной лингвистике и ее приложениям, 30 мая–5 июня 2001 г. Т. 2.— Аксаково, 2001.— С. 273–279.

9. Химическая энциклопедия: В 5 т. Т. 3. Меди — Полимерные.— М.: Большая Российская энцикл., 1992.— С. 290–293.

10. Зацман И. М. Логико-семантические модели полнотекстовых научных документов // НТИ. Сер. 2.— 1999.— No 5.

11. Лунев Б. С., Наумова О. Б. Атлас форм рельефа: В 2 т. Т. 1. Основные рельефообразующие факторы Земли.— Пермь: Пермский ун-т, 1998.— 296 с.

12. Zatsman I. M. Semantic Encoding and Markup of Georeferenced Documents in Polythematic Digital Libraries of Scientific Literature // Third All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Petrozavodsk, September 11–13, 2001 г.— Petrozavodsk: KarRC RAS, 2001.— P. 136–142.

13. Geography Markup Language (GML) 1.0. OGC Request 11: A Request for Comments: OpenGIS Geography Markup Language Specification (URL: http://www.opengis.org/techno/RFC11.bak/GMLRFCV1_0.html).

Материал поступил в редакцию 12.02.01.