ПРИНЦИПЫ СЕМАНТИЧЕСКОГО КОДИРОВАНИЯ ПЕРВИЧНЫХ ГЕОДАННЫХ [I)]

И.М. Зацман, И.В. Землянов

ИПИ РАН, ГОИН, Москва

Рассматриваются принципы создания вербально-образного геотезауруса, предназначенного для представления знаний о классификации форм рельефа и элементов гидрографической сети в виде дескрипторов и связей между ними. Включение в геотезаурус образных дескрипторов позволяет решать задачи семантического кодирования геоизображений и других первичных образных геоданных, например, электронных карт, и концептуального поиска геоизображений в электронных геобиблиотеках.

Введение

Организация концептуального поиска в электронных геобиблиотеках зависит от ряда факторов, определяющих и степень ее сложности, и постановки конкретных задач концептуального поиска. Принципиальным отличием геобиблиотек является хранение в них электронных карт и других геоизображений, содержащих информацию о широком спектре геообъектов, например, форм рельефа и элементов гидрографической сети.

Необходимость поиска геоизображений по их содержанию требует определить те элементарные единицы, которые могут быть минимальными носителями содержания геоизображений. Для вербальных текстов минимальной единицей смысла, связанной с вербальной формой его представления, является вербальный знак. Возможность членения вербального текста на формы знаков, а форм на знаковые примитивы – принцип двойного членения – является основой для разработки методов поиска текстов в электронных библиотеках вербальных документов. Для геоизображений, которые не являются линейной дискретной последовательностью форм знаков и знаковых примитивов, принцип двойного членения в общем случае не действует. Более того, в настоящее время отсутствует общепринятое понятие образного знака для изображений.

Основное содержание доклада заключается в описании принципов создания вербально-образного геотезауруса, которые рассматриваются во взаимосвязи с определением понятия образного знака.

Вербальные и образные знаки

Согласно теории семиосферы Ю.М. Лотмана в случае вербальных текстов основным носителем значения является вербальный знак, а цепочка форм знаков, составляющая текст – вторична. Значение ее производно от значений знаков. Для многих видов невербальных текстов он (текст) первичен. Ю.М. Лотман пишет: «Он является носителем основного значения. По своей природе, он не дискретен, а континуален. Смысл его не организуется ни линейной, ни временной последовательностью, а "размазан" в n-мерном семантическом пространстве данного текста (полотна картины, сцены, экрана, ритуального действа, общественного поведения или сна). В текстах этого типа именно текст является носителем значения. Выделение составляющих его знаков бывает затруднительно и порой носит искусственный характер» [1, с. 178].

В работах [2; 3] показано, что носителями содержания образных текстов, хранимых в электронных библиотеках, могут являться те знаки, которые отсутствуют в этих текстах. Приведены примеры географических карт, содержание которых частично можно передать образными знаками, отсутствующими в картах[II)]. Карты и другие геоизображения рассматриваются как один из широко распространенных видов образных текстов. В этих работах предложено дополнить приведенные положения теории семиосферы применительно к образным текстам следующим образом: «носителями содержания образных текстов могут являться те знаки, формы которых отсутствуют в этих текстах, но при этом значения таких знаков и их сочетаний отражают содержание образных текстов».

Предлагаемое дополнение позволяет не ограничиваться только выделением форм знаков в образных текстах, которое бывает затруднительно и порой носит искусственный характер. Для некоторых видов невербальных текстов, например образных, имеется возможность искусственным путем определить образные знаки, формы которых отсутствуют в текстах, но при этом значения этих знаков передают содержание образных текстов, например, геоизображений.

[I)] Работа выполнена при частичной поддержке РФФИ, грант № 03-05-65264.

[II)] Аналогичную ситуацию иногда можно наблюдать и при обработке вербальных текстов, например при выборе ключевых слов для научных статей в тех случаях, когда их содержание передается ключевыми словами, отсутствующими в статьях.

Примеры, иллюстрирующие подобную возможность, рассматриваются в работе [3]. Искусственность определения образных знаков говорит также о вторичности и условности формы по сравнению со значением образного знака.

Рассмотренные положения теории семиосферы Ю.М. Лотмана играют центральную роль в постановке и решении задач концептуального поиска изображений. Из них следует, например, что при организации поиска изображений в пределах электронных геобиблиотек и других геоинформационных систем нужно искать более устойчивые «маркеры» значений, чем формы знаков. В качестве подобных «маркеров» могут выступать коды элементарных концептов, примеры которых рассматриваются далее.

Для некоторых предметных областей образное представление концептов является характерной экспликацией знаний. В науках о Земле само существование географических карт и других геоизображений свидетельствует о недостаточности вербальных средств представления знаний о Земле.

В работе [4] дано описание эксперимента по исследованию информационных модальностей, используемых пользователями информационных систем для описания пространственных объектов в запросах. В процессе эксперимента было зафиксировано, что они пользовались вербальной и образной модальностями для описания географических концептов. В результате этого эксперимента более половины пространственных объектов были представлены в обеих информационных модальностях. Эксперимент продемонстрировал, что одной информационной модальности – вербальной или образной – может оказаться недостаточно при построении пользователями нужных им запросов.

Для процесса кодирования геоизображений с использованием образных знаков можно провести аналогию с тем процессом, в котором вербальные знаки используются для индексирования и поиска вербальных компонентов документов. Но есть в постановке этой задачи и принципиальное отличие от вербального индексирования. Это отличие заключается в том, что заранее, до кодирования геоизображений, образные знаки должны быть определены в явной форме на основе относительно устойчивой системы семантических отношений. В используемой системе форма знака должна репрезентировать конвенционально приданное ему значение.

Образные знаки в электронных геобиблиотеках

Определение системы образных знаков, применяемых для кодирования геоизображений, является центральным моментом доклада. Основные принципы построения систем образных знаков были определены в рамках концепции вербально-образного представления знаний в электронных библиотеках [5]. Рассмотрим пять основных положений этой концепции, используемые в процессе определения системы образных знаков.

Основная идея концепции вербально-образного представления знаний заключается в том, что система образных знаков строится заранее в одной или нескольких электронных библиотеках как относительно замкнутых информационных системах. Это **первое положение** концепции, отличающее предлагаемый подход от вербального индексирования и поиска. В последнем случае используются системы парадигматических, синтагматических и семантических отношений, которые являются результатом изучения существующих естественно-языковых систем, а не результатом их построения.

Второе положение концепции вербально-образного представления знаний заключается в разграничении сфер использования следующих процессов:

1) процесс построения (определения) системы образных знаков для электронной библиотеки,
2) процесс вычленения образных знаков, определенных в рамках первого процесса, в образных компонентах документов электронной библиотеки;
3) конструирование образных компонентов документов из системы знаков, определенных в рамках первого процесса.

В задаче кодирования предполагается, что из трех перечисленных процессов определение системы образных знаков электронной библиотеки охватывает максимальное число образных компонентов документов в электронной геобиблиотеке с точки зрения передачи их смысла. Знаки построенной системы предназначены для отражения содержания максимального числа геоизображений электронной библиотеки. Содержание каждого геоизображения может быть только частично передано сочетанием образных знаков. При этом не требуется, чтобы формы знаков, передающих содержание геоизображений, обязательно в них присутствовали.

Следующим в порядке уменьшения сферы использования является процесс вычленения образных знаков. Среди всего массива геоизображений выделяется подмножество, содержание элементов которых может быть только частично отражено образными знаками электронной библиотеки, определенными в рамках первого процесса. При этом из каждого геоизображения этого подмножества можно вычленить те знаки, которые передают содержание этих геоизображений. Однако из них нельзя сконструировать эти компоненты полностью, так как выделенные знаки и их сочетания не отражают все содержательные аспекты геоизображений этого подмножества.

И еще меньшую сферу использования имеет процесс конструирования геоизображений из знаков. Другими словами, теоретически допускается

существование таких геоизображений, содержание которых может быть полностью выражено конечным числом образных знаков электронной библиотеки, вычленяемых из этий геоизображений. Кроме того, из них можно сконструировать эти геоизображения полностью, так как выделенные знаки и их сочетания полностью отражают все их содержательные аспекты.

Третье положение концепции вербально-образного представления знаний заключается в том, что в качестве индексов образных компонентов документов электронной библиотеки, кроме вербальных знаков, используются образные и вербально-образные знаки, которые по определению являются дескрипторами вербально-образного тезауруса.

Четвертое положение концепции касается роли графических примитивов. Если при индексировании и поиске образных компонентов используются графические примитивы, то только в составе образных знаков.

Пятое положение концепции говорит о том, что в качестве конвенциональной основы построения системы образных знаков используются существующие в науках, областях знаний и научных специальностях конвенциональные вербально-образные системы классификаций объектов и явлений. Для геоизображений основой построения системы образных знаков является набор общих понятий или концептов системы классификации геообъектов и явлений, выбранной для электронной библиотеки.

В этом случае общность конвенциональных основ вербальных и образных знаковых систем заключается в использовании системы семантических отношений. Для вербальной знаковой системы конвенциональной основой является система семантических отношений естественного языка, а для образной знаковой системы электронной библиотеки научных документов – система семантических отношений выбранной научной классификации.

Пример построения образных знаков

В качестве демонстрационного примера рассмотрим классификацию устьевых областей рек и их частей по морфологическим признакам из работы [6]. В соответствии с первым положением концепции система образных знаков строится как искусственная система, что является ключевым ее отличием от вербальных знаковых систем.

В классификации элементов гидрографической сети, при последовательном переходе от более абстрактных понятий (элемент гидрографической сети) к более конкретным географическим объектам, на одном из уровней появляется понятие «устьевая область реки», которому соответствуют географические объекты, охватывающие район впадения реки в приемный водоем (океан, море, озеро). Ключевой элемент в построении образных знаков представляет собой таблицу, построенную на основе классификации устьевых областей рек и их частей – устьевого участка и устьевого взморья – по морфологическим признакам (см. табл. 1).

В соответствии с третьим положением концепции вербально-образного представления знаний, в качестве индексов образных компонентов документов электронной библиотеки, кроме вербальных знаков, используются также образные и вербально-образные знаки, являющиеся по определению дескрипторами тезауруса электронной геобиблиотеки.

Будем далее использовать информацию этой таблицы для построения дескрипторов вербально-образного тезауруса электронной библиотеки. Перечисленные в третьей колонке таблицы названия устьевых областей в работе [6] сопровождаются стилизованными изображениями – пиктограммами, которые будем использовать как формы образных дескрипторов вербально-образного тезауруса электронной библиотеки.

Определим систему и правила назначения кодов дескрипторам устьевых областей рек, их формам и значениям (элементарным концептам).

Для кодирования шести устьевых областей рек из табл. 1 в любых геоизображениях, содержащих элементы гидрографической сети, построим серию вербальных и образных дескрипторов с шестью значениями. Их значениям присвоим коды 001к–110к (см. табл. 2) в соответствии с третьим столбцом табл. 1, начиная с устьевой области «Простая без блокирующей косы» (001к) и кончая – «Дельтовая с дельтой выдвижения» (110к). Значения, объединенные с образными формами в виде пиктограмм (см. рис. 1), будем называть образными дескрипторами тезауруса электронной геобиблиотеки. Обозначим их коды как 001до–110до.

Значения, которые объединены с соответствующими им кодами 001к–110к, являются примерами семокодов. Пиктограммы как образные формы, которые объединены с соответствующими им кодами 001о–110о, являются примерами образных формокодов. Понятия семокода и формокода определены в работе [7].

Шесть значений с кодами 001к–110к, которые объединены с названиями на русском языке, будем называть вербальными дескрипторами русскоязычной части тезауруса. Обозначим их коды как 001др–110др. Родовое понятие для них выражается дескриптором «устьевая область реки». Названия на русском языке, которые объединены с соответствующими им кодами 001р–110р, являются примерами вербальных формокодов.

Шесть значений с кодами 001к–110к, которые объединены с названиями на английском языке, будем называть вербальными дескрипторами англоязычной части тезауруса. Обозначим их коды как 001да–110да. Родовое понятие для них выражается дескриптором «river mouth».

Названия устьевых областей рек и их стилизованные изображения из системы классификации использовались для построения вербальных и образных дескрипторов тезауруса в соответствии с пятым положением концепции. Отношения между дескрипторами наследуют отношения между этими геообъектами в классификации.

Устьевой участок реки	Устьевое взморье	Устьевая область реки
Однорукавный (бездельтовый)	Открытое без блокирующей косы	Простая без блокирующей косы
	Полузакрытое – без блокирующей косы – с блокирующей косой	Эстуарная – без блокирующей косы – с блокирующей косой
Мало- и много-рукавный (дельтовый)	Полузакрытое – без блокирующей косы – с блокирующей косой	Эстуарно-дельтовая – без блокирующей косы – с блокирующей косой
	Открытое	Дельтовая с дельтой выдвижения

Табл. 1. *Классификация устьевых областей рек и их частей по морфологическим признакам [6]*

Коды значений вербальных и образных дескрипторов, которые по определению совпадают	Названия устьевых областей рек на русском и английском языках и коды названий (в скобках)	Номер рисунка и коды пиктограмм (форм образных дескрипторов)
001кр≡001ка≡ 001ко≡001к	Простая без блокирующей косы (001р) – Simple river mouth without blocked spit (001a)	Рис. 1, код 001о
010кр≡010ка≡ 010ко≡010к	Эстуарная без блокирующей косы (010р) – Estuary without blocked spit (010a)	Рис. 1, код 010о
011кр≡011ка≡ 011ко≡011к	Эстуарная с блокирующей косой (011р) – Estuary with blocked spit (011a)	Рис. 1, код 011о
100кр≡100ка≡ 100ко≡100к	Эстуарно-дельтовая без блокирующей косы (100р) – Silt delta without blocked spit (100a)	Рис. 1, код 100о
101кр≡101ка≡ 101ко≡101к	Эстуарно-дельтовая с блокирующей косой (101р) – Silt delta with blocked spit (101a)	Рис. 1, код 101о
110кр≡110ка≡ 110ко≡110к	Дельтовая с дельтой выдвижения (110р) – Protruding delta (110a)	Рис. 1, код 110о

Табл. 2*. Коды дескрипторов вербально-образного тезауруса, их форм и значений*

В таблице использованы следующие обозначения:

- 001к*–110к* – коды значений дескрипторов, назначенные шести понятиям, определенным в рамках выбранной таксономии устьевых областей («к» – от слова «концепт», * – обозначает литеры «р», «а» или «о»);
- 001р–110р – коды названий видов устьевых областей рек на русском языке;
- 001а–110а – коды названий видов устьевых областей на английском языке;
- 001о–110о – коды форм образных дескрипторов, отражающих в графической форме особенности морфологического строения того или иного вида устьевой области реки.

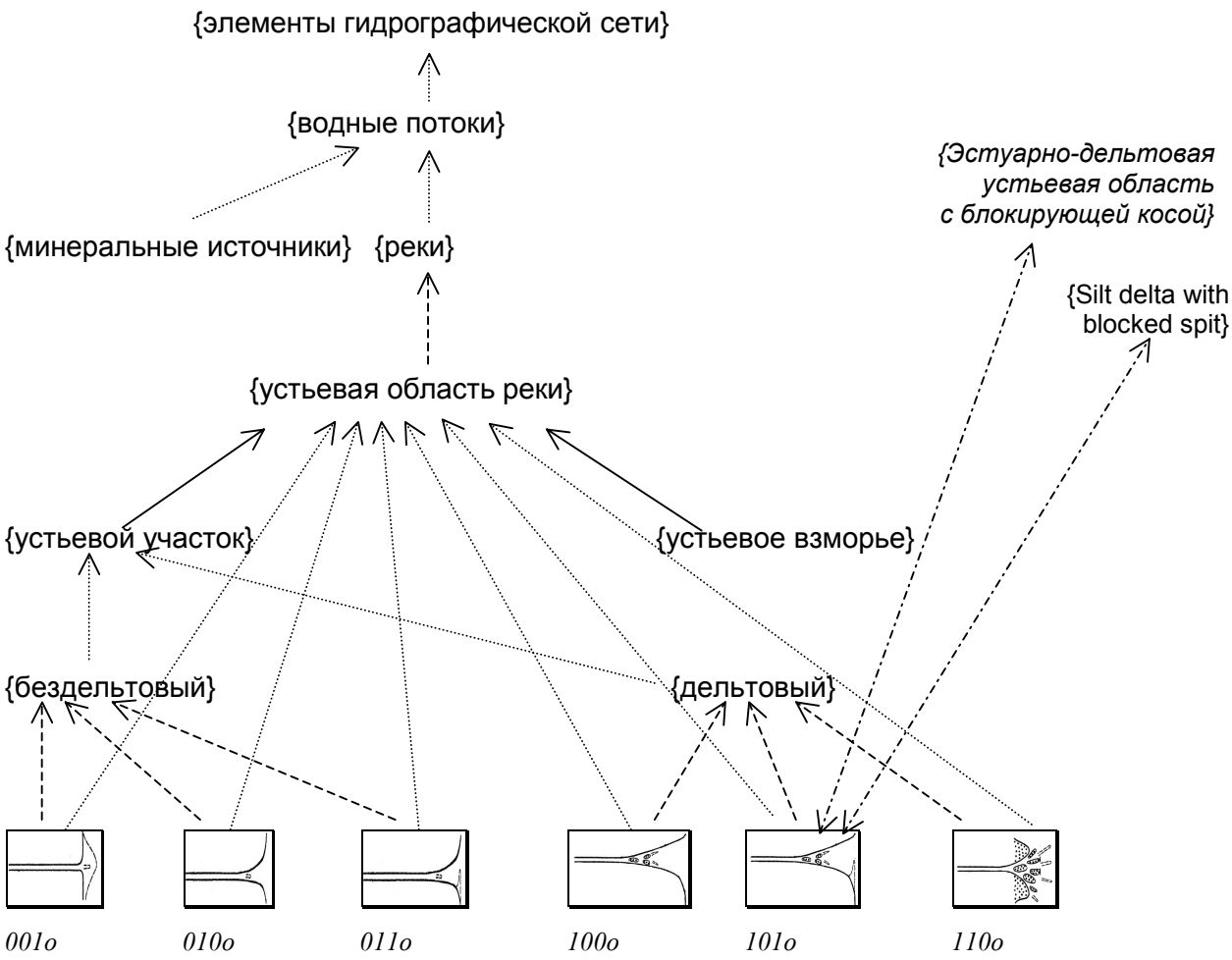

Рис. 1. *Фрагмент вербально-образного тезауруса*

Фрагмент вербально-образного тезауруса, включающий формы дескрипторов и отношения между ними, приведен на рис. 1. Точечными стрелками на рисунке условно обозначены родовидовые отношения, сплошными – отношения «часть–целое», штриховыми – ассоциативные отношения.

Не приведены коды понятий и названий устьевых областей на русском и английском языках. Приведены только условные алфавитно-цифровые обозначения кодов форм образных дескрипторов устьевых областей рек в рамках выбранной системы их классификации. Виды устьевого взморья и минеральных источников не показаны.

Объединяя второе, третье и пятое положения концепции, в итоге получаем следующую последовательность шагов от выбранной классификации до знаков образных компонентов документов электронной библиотеки: научная система классификации – дескрипторы вербально-образного тезауруса – образные знаки электронной библиотеки.

В соответствии со вторым положением образные знаки в этой задаче строятся только на основе системы классификации в рамках первого процесса, т.е. без обязательного вычленения форм знаков из образных компонентов документов.

Важно отметить, что использование стилизованных изображений позволяет устранить языковой барьер, возникающий при использовании естественных языков. Кроме того, в рассматриваемой задаче стилизованные изображения, которые предлагается использовать в качестве образных дескрипторов геоизображений документов электронной геобиблиотеки, в графической форме отражают особенности морфологического строения того или иного вида устьевой области.

В кодировании геоизображений будем использовать новый вид связи между дескрипторами вербально-образного тезауруса, который не встречается в вербальных тезаурусах. В работе [5] этот вид связи было предложено называть отношением семиотической или гетеромодальной синонимии. Если значения вербального и образного дескриптора совпадают и это совпадение отражено в вербально-образном тезаурусе, то будем говорить, что эти дескрипторы являются семиотическими или гетеромодальными синонимами.

В рассматриваемой задаче название и стилизованное изображение каждого вида устьевой области реки являются формами вербального и образного знаков. Эти знаки связаны отношением семиотической синонимии, так как их значения в

тезаурусе совпадают. Например, образный дескриптор, форма которого изображена на рис. 1 (код 101о), имеет два семиотических синонима – вербальные дескрипторы с названиями «Эстуарно-дельтовая устьевая область с блокирующей косой» и «Silt delta with blocked spit». Отношения семиотической синонимии на рисунке обозначены штрих-пунктирными стрелками.

Индексирование геоизображений

В предыдущем параграфе дано описание фрагмента вербально-образного тезауруса, дескрипторы которого предлагается использовать при индексировании геоизображений. Как следует из табл. 2, кроме образных дескрипторов, на основе выбранной системы классификации были определены и вербальные дескрипторы.

В описании задачи индексирования геоизображений используется понятие «вербально-образный тезаурус», определенное в работе [5] при описании концепции вербально-образного представления знаний в электронных библиотеках и других информационных системах. В рамках этой концепции только после построения и включения в тезаурус необходимых дескрипторов появляется возможность образного индексирования геоизображений документов электронной геобиблиотеки. Это связано с тем, что в процессах такого индексирования геоизображений можно использовать только дескрипторы.

В концепции образные знаки документов электронной геобиблиотеки трактуются как синонимы образных дескрипторов. В случае вербально-образного тезауруса электронной геобиблиотеки его дескрипторы строятся на основе следующих положений концепции:

- образные дескрипторы могут быть и мотивированными, и немотивированными;
- дескрипторы имеют одно значение в системе отношений тезауруса;
- значения образных дескрипторов тезауруса и, следовательно, образных знаков документов электронной библиотеки определяются на основе систем семантических отношений отобранных научных классификаций объектов и явлений;
- отношения между дескрипторами включают традиционные для вербальных тезаурусов системы отношений;
- отношения между дескрипторами могут также включать новые виды отношений, которые в вербальных тезаурусах не встречаются (например, отношения семиотической или гетеромодальной синонимии).

Для документов по наукам о Земле речь будет идти о конвенциональных научных классификациях геообъектов и явлений. С учетом перечисленных положений концепции, индексирование каждого геоизображения определим как установление соответствия между ним и конечным числом отобранных из тезауруса вербальных, образных и вербально-образных дескрипторов, которое обладает следующими свойствами:

- содержательные аспекты геоизображения должны быть выражены отобранными дескрипторами и семантическими отношениями между ними полностью или частично;
- если отобранный дескриптор является образным, то он может не совпадать ни с одним из фрагментов индексируемого геоизображения;
- в геоизображении могут существовать фрагменты, содержательные аспекты которых не отражены отобранными дескрипторами;
- отношения между дескрипторами геоизображения могут не отражать всю полноту семантических отношений, которую можно наблюдать в геоизображении;
- если дескриптор отобран для индексирования геоизображения и имеет онтологические пары в тезаурусе, то автоматически они также становятся индексами этого геоизображения.

Семантические отношения между отобранными дескрипторами могут быть двух видов. Отношения первого рода наследуются из системы отношений тезауруса, т.е. являются универсальными в пределах электронной библиотеки. Отношения второго рода формируются в процессе декомпозиции и построения логико-семантических моделей документов электронной геобиблиотеки [5].

Деление на отношения первого и второго родов является достаточно условным. Возможен случай, когда при развитии тезауруса, отношения второго рода, полученные при обработке нескольких геоизображений, отражают устойчивые связи между объектами и явлениями в науках о Земле, что позволяет включить их в систему отношений вербально-образного тезауруса.

Практическое применение

Подход, изложенный в предыдущих разделах, в настоящее время реализуется в рамках проекта «Электронный атлас устьев рек» [8]. Предметом проекта является создание электронного учебно-научного атласа морских устьев рек Евразии. Основой для создания атласа служат спутниковые изображения устьевых областей из космоса, представленные с различным пространственным разрешением и в разных масштабах, объединенные с фрагментами физико-географических и тематических карт для соответствующих районов и текстовыми описаниями. Совместное представление содержания физико-географических и тематических карт, текстовых описаний конкретных устьевых областей и спутниковых изображений в различных спектральных диапазонах используется сегодня для решения следующего круга задач:

1) классификация морских устьев рек с использованием космических снимков,

фрагментов географических карт и текстовых описаний;
2) изучение особенностей строения морских устьев рек различных типов с использованием геоизображений;
3) изучение морфологии и закономерностей формирования основных структурных элементов морских устьев рек и основных гидролого-морфологических процессов в устьях рек на основе имеющихся геоизображений;
4) анализ структуры морских устьев рек и протекающих в них гидролого-морфологических процессов с использованием синтезированных спутниковых многозональных геоизображений.

Объем и разнообразие графического информационного материала, представляемого в атласе, уже сейчас позволяет говорить об актуальности создания системы навигации и поиска геоизображений, основанной на использовании вербально-образного тезауруса.

Основные выводы

Использование образных знаков для кодирования геоизображений отличается от использования вербальных знаков. Это отличие заключается в том, что в последнем случае используются системы парадигматических, синтагматических и семантических отношений, которые являются результатом изучения существующих естественно-языковых знаковых систем, а не результатом их построения.

Образные знаки определяются на основе относительно устойчивой системы семантических отношений – научной классификации. В такой системе форма образного знака репрезентирует конвенционально приданное ему значение.

Использование образных дескрипторов для кодирования карт и их фрагментов дает возможность многократного использования уже имеющихся результатов индексирования. Например, если назначены образные дескрипторы для одной топографической карты некоторого масштаба, то для каждой последующей карты этой местности с тем же масштабом необходимо в процессе индексирования обрабатывать только ее отличия от первоначально проиндексированной карты.

Список литературы:

1. Лотман Ю.М. Семиосфера. СПб.: «Искусство-СПб», 2000.
2. Зацман И.М. Визуально-мотивированное представление знаний в электронных библиотеках научных документов // Труды 4-й Всероссийской конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (Дубна, 15–17 октября 2002г.): В 2 т. Т. 1. Дубна: ОИЯИ, 2002. С. 120–135.
3. Zatsman I. Pictorial Signs for Geoimages in Digital Libraries // European Journal for Semiotic Studies. Vol. 15. N. 2–4. 2003. P. 609–620.
4. Schlaisich I., Egenhofer M. Multimodal Spatial Querying: What People Sketch and Talk About // In Proceedings of the 1st International Conference on Universal Access in Human-Computer Interaction, August 2001. New Orleans, LA. / Ed. by C. Stephanidis. New Orleans: 2001. P. 732–736.
5. Зацман И.М. Концептуальный поиск и качество информации. М.: Наука, 2003.
6. Михайлов В.Н. Гидрология устьев рек: Методическое пособие. М.: Изд-во МГУ, 1996.
7. Зацман И.М. Семиотический анализ человеко-машинного взаимодействия в технологиях поиска (см. наст. сборник докладов).
8. Землянов И.В., Горелиц О.В. Электронный атлас морских устьев рек России. Сборник докладов Первого Общероссийского научно-практического семинара "Электронная Земля. Электронная Россия. Электронная Москва: методология и технологии". М.: ИПИ РАН, 2002. С. 81-84.