

Семантическое, информационное и знаковое кодирование патентных документов электронных библиотек

© И.М. Зацман

Институт проблем информатики РАН
imz@al70.ipi.ac.ru

Аннотация

Доклад посвящен развитию схемы концептуального поиска в электронных библиотеках патентных документов. В качестве исходных данных для доклада взята постановка проблемы концептуального поиска (ПКП) в электронных библиотеках, предложенная в монографии [1], которая содержит описание двух вариантов схемы концептуального поиска. Варианты отличаются источниками информационных ресурсов, которые используются для формирования электронных библиотек, а именно - человек и/или техническая система [2]. Основная цель доклада заключается в определении процедуры комплексного кодирования документов, в рамках которой различаются семантический, информационный и знаковый виды кодирования, что является принципиальным развитием ранее разработанных вариантов схемы концептуального поиска.

1 Введение

На конференции RCDL2004 в основном докладе Е.А. Fox [3] изложен подход, обозначенный аббревиатурой 5S (Societies, Scenarios, Spaces, Structures, and Streams). Для каждой из пяти перечисленных составляющих определены модели, включая их примитивы, формализм и назначение. При этом потоки (Streams) были определены как последовательности элементов произвольных типов.

Однако если патентные документы содержат образные континуальные компоненты, например, изображения, то их представление в виде последовательностей семантически значимых или смысловых элементов является самостоятельной и непростой задачей. Конечно, для целей хранения мы можем представить такие компоненты в растровой форме и кодировать в электронных библиотеках каждую точку растра с

помощью битовых структур. Однако для поиска хотелось бы кодировать также содержание изображений с помощью нулей и единиц цифровой среды, чтобы затем искать изображения по их содержанию, а не только по графическим примитивам.

В качестве одного из возможных подходов к решению этой проблемы в докладе предлагается ввести процедуру комплексного кодирования, в рамках которой различаются три следующих вида кодирования документов:

- семантическое (кодирование концептов документов),
- информационное (кодирование форм представления концептов, которые будем называть информационными объектами),
- знаковое (кодирование сочетания концепта и одной из форм его представления, которое в семиотике традиционно называется знаком).

Предположим, что некоторый концепт, как ментальную единицу знаний человека, можно выразить словами Солнце, Sun или пиктограммой, например ☀. Если код концепта во всех трех случаях будет один и тот же, но будет отличаться от кода фирмы Sun, то тогда с помощью кодов этих концептов в англоязычных документах электронной библиотеки можно будет достаточно просто отличать солнце от упоминаний фирмы Sun в случае абсолютно одинакового написания этого слова с прописной буквы.

Приведенный пример наглядно демонстрирует актуальность решения проблемы комплексного кодирования вербальных и невербальных компонентов документов с использованием кодов концептов, информационных объектов и знаков. Отметим, что для вербальных компонентов семантическое кодирование традиционно выполняется в процессе индексирования с использованием вербальных тезаурусов [4].

2 Пример предметной области

В докладе приводится краткое описание постановки ПКП для вербальных и невербальных форм представления искомых концептов, в том числе и для случая, когда пользователь не знает, в какой информационной модальности (текст, карта, рисунок, таблица, график, формула или их сочетание) выражен интересующий его концепт

(концепты). В постановке ПКП предполагается, что пользователь может использовать любую из допустимых в электронной библиотеке информационных модальностей для представления искомого концепта в запросе. При этом форма выражения искомого концепта в документе заранее пользователю может быть неизвестна, но он должен знать тот концепт, формы представления которого он ищет.

Актуальной предметной областью, в которой форма выражения искомого концепта заранее пользователю может быть неизвестна, является патентное дело. Цель патентного поиска заключается в том, чтобы для новой заявки на изобретение найти в патентном фонде уже имеющиеся описания патентов, содержащие описание того же по смыслу решения. Если при поиске такие патенты будут найдены, то это может служить основанием для отклонения поступившей новой заявки. Однако в имеющихся патентах решение может быть описано словами, а в новой заявке то же решение может быть выражено в виде формул или рисунков с формулами. Если в технологии формирования патентной электронной библиотеки, способах ее ведения и методах поиска удалось бы учитывать при обработке запросов возможные формы представления искомого

концептов, то описания существующих решений могли бы быть найдены независимо от формы их представления – с помощью слов, формул или рисунков.

Если в предметной области электронной библиотеки существуют концепты, которые допускают вербальные и невербальные формы их представления, то постановка ПКП значительно усложняется, так как модальности форм представления концептов пользователь электронной библиотеки в общем случае заранее не знает.

В этом случае, комплексное кодирование патентных документов позволяет сформулировать задачи концептуального поиска и предложить подход к их решению, основанный на комплексном кодировании патентных документов, включая семантическое кодирование.

3 Схема концептуального поиска

В работе [1] рассмотрены достаточно подробно два варианта схемы концептуального поиска научных документов в электронных библиотеках. Первый вариант схемы, состоящий из 8 процессов, приведен на рис. 1. На рисунке с помощью штриховой и штрих-пунктирной линий выделены три уровня схемы (верхний, средний и нижний).



Рис. 1. Схема концептуального поиска (первый вариант)

Уровни выделены согласно трем терминам концептуального поиска «знания», «информация» и «цифровые коды», которые по определению могут использоваться только на соответствующих уровнях.

В докладе термин «информация» используется для обозначения знаковой информации. Другие рассматриваемые виды информации обозначены терминами «цифровые коды», «данные» и «цифровые данные», что дает возможность назвать и точно соотнести каждый вид информации и уровень схемы.

Второй вариант схемы приведен на рис. 2. Числами от 1 до 12 на нем обозначены 12 процессов второго варианта схемы концептуального поиска. На рисунке с помощью линий выделены пять уровней схемы согласно пяти терминам концептуального поиска «знания», «информация», «цифровые коды», «данные» и «цифровые данные», которые по определению могут использоваться только на соответствующих уровнях.

Добавление двух терминов «данные» и «цифровые данные» увеличивает перечень процессов во втором варианте с 8 до 12.



Рис. 2. Схема концептуального поиска (второй вариант)

Верхний уровень на рис. 2 практически совпадает с верхним уровнем на рис. 1. Отличие заключается только в том, что в явной форме не обозначены авторские и пользовательские концепты – оставлены только интерпретированный и искомый концепты. Нижний уровень кодов и уровень цифровых данных укорочены слева, чтобы условно изобразить границу между средним уровнем информации и уровнем данных в виде точечной линии.

Приведенные варианты схемы концептуального поиска содержат названия основных процессов, показывают связи между ними, а также отношения между процессами и уровнями схемы. О методах реализации процессов эти варианты нам

ничего не сообщают. Однако приведенные варианты схемы помогают наглядно зафиксировать место действия каждого процесса концептуального поиска.

4 Процессы схемы концептуального поиска

Основной целью доклада является развитие схемы концептуального поиска документов в электронных библиотеках с использованием тезауруса. Важно отметить, что развитие схемы должно учитывать действие объективного закона асимметрии знаковых систем [5] для следующих процессов:

- представление знаний (концептов) в документах (процесс 1 на рис. 1 и 2),
- информационное представление знаний пользователей в виде запроса (формы знаков) на проведение поиска в электронной библиотеке (процесс 4 на рис. 1 и 2),
- понимание (семантическая интерпретация) пользователем найденных документов (процесс 8 на рис. 1 и 2) как соотнесение форм знаков и их значений.

По определению понятия семиотического знака его значение и форма относятся, соответственно, к верхнему и среднему уровням в обоих вариантах схемы. Далее будем рассматривать только первый вариант схемы, так как процессы 1, 4 и 8 одинаковы для обоих вариантов.

Все три процесса связаны с переходами между верхним и средним уровнями схемы. Переходы от верхнего уровня к среднему для процессов 1 и 4 условно обозначены двумя прямоугольниками «1. Создание авторских документов» и «4. Формирование информационных объектов запросов».

Эти прямоугольники изображены на среднем уровне, к которому относятся результаты процессов 1 и 4. Однако основные объекты процессов 1 и 4 – концепты авторов и

пользователей, информационные объекты документов и запросов – относятся к разным уровням схемы: «знания» и «информация».

Обратный переход со среднего уровня на верхний условно обозначен прямоугольником «8. Интерпретация пользователем авторских информационных объектов», т.е. семантическая интерпретация на основе восстановленных информационных объектов в его понимании.

Этот прямоугольник на схеме изображен на верхнем уровне, к которому относятся результаты процесса семантической интерпретации вербальных и/или невербальных информационных объектов найденных документов.

Основные объекты процесса 8 – концепты пользователя и информационные объекты найденных документов – также относятся к разным уровням схемы: «знания» и «информация». Рассмотрим отношения объектов верхнего и среднего уровней с точки зрения семиотики, используя рис. 3. Он создан на основе рис. 1, у которого оставлены три уровня схемы и границы между ними. По сравнению с рис. 1 удалены восемь процессов схемы и добавлена граница между верхним и нижним уровнями, которая условно обозначена сплошной двойной линией.

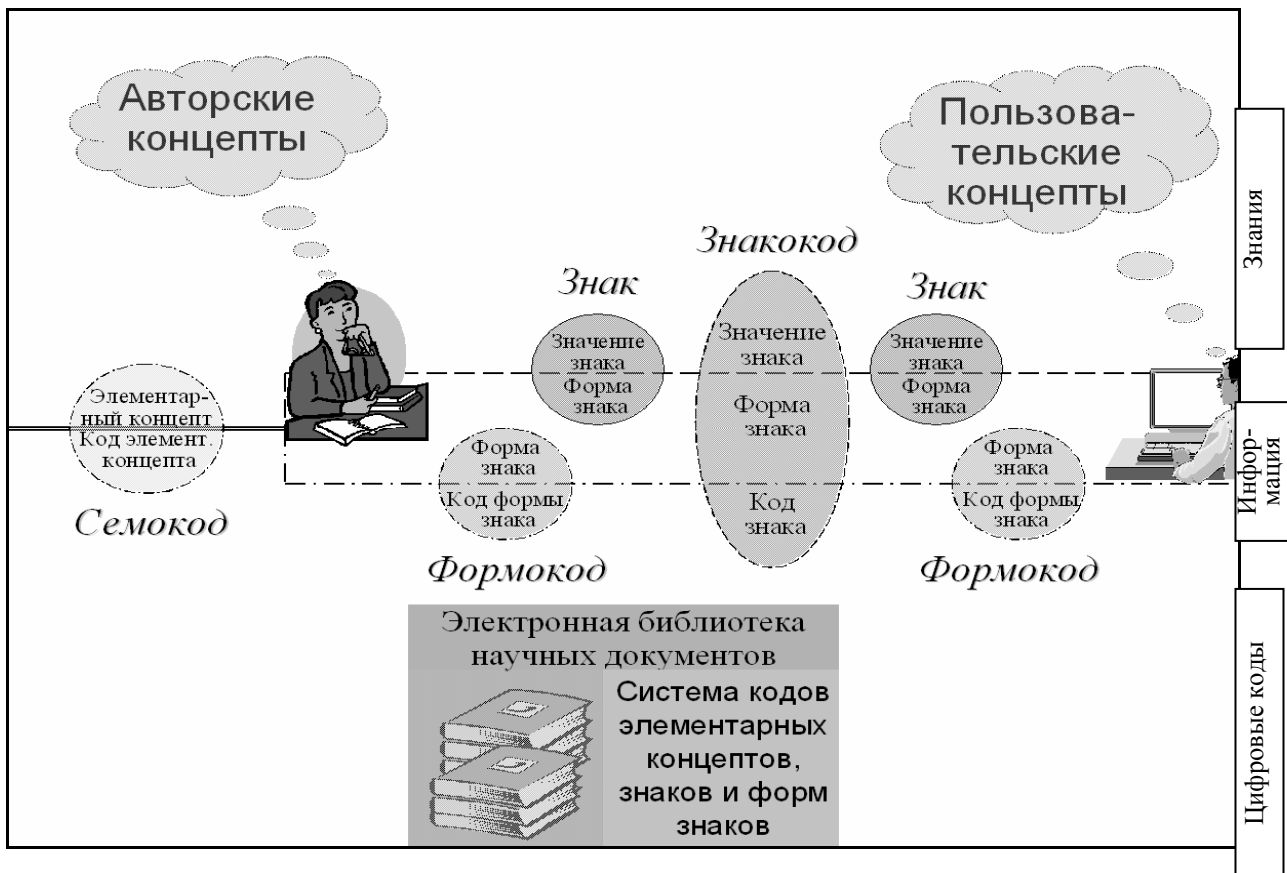


Рис. 3. Знаки, формокоды, семокоды и знакокоды

Уточнено также, что составной частью электронной библиотеки является система кодов

элементарных концептов, знаков и форм знаков. На рис. 1 была обозначена система кодов, но без

перечисления отдельных их категорий. К первой категории по определению отнесем коды знаков, ко второй – коды их форм, а к третьей – коды их значений или элементарных концептов. На рис. 3 указаны все три категории кодов.

Знаки на этом рисунке условно обозначены кружками на границе между верхним и средним уровнями схемы. Именно на этой границе действует закон асимметрии знаковых систем. Семиотическое понятие знака как двуединой сущности, рассмотренное в контексте схемы концептуального поиска, привлекает внимание к границе между средним и нижним уровнями.

В работе [1] по аналогии со знаком было определено новое понятие, интегрирующее форму знака и код этой формы, которое называется **формокодом**. Если использовать аналогию со знаками как элементарными единицами конвенционального представления знаний в среде социальных коммуникаций, то формокоды являются элементарными единицами конвенционального представления информации в цифровой среде. Было также определено еще одно новое понятие, интегрирующее элементарный концепт и его код, которое называется **семокодом**.

Важно отметить, что процессы индексирования и поиска документов относятся к цифровой среде, т.е. к нижнему уровню схемы «цифровые коды». Попытаемся изменить схему таким образом, чтобы обеспечить взаимно однозначные отношения кодов и значений знаков (концептов) в схеме, используя понятие семокода. В тех случаях, когда выбор авторами формы представления концепта не влияет на его код в цифровой среде, исключается трансляция закона асимметрии знаковых систем в цифровую среду.

Другими словами, если код концепта в электронной библиотеке не зависит от формы его представления на среднем уровне, то действие закона асимметрии будет ограничиваться только средним уровнем и не будет транслироваться на нижний уровень, к которому относятся процессы индексирования и поиска документов.

Отсюда следует искомое направление развития схемы – выбор формы представления концептов в документах электронной библиотеки не должен влиять на коды концептов автора, принадлежащих цифровой среде. Но реально ли хотя бы частично ограничить действие закона асимметрии средним уровнем, если научные документы и патенты на экране компьютера и на бумаге являются только формами вербальных и невербальных знаков или их сочетаниями, а формы по определению принадлежат к среднему уровню схемы?

Прежде чем ответить на этот вопрос, рассмотрим пример укрупненной технологической схемы подготовки электронных форм патентных заявок для формирования патентных библиотек (рис. 4). Отметим, что описания изобретений могут включать математические и структурные

химические формулы, графики, таблицы, диаграммы, чертежи и другие невербальные информационные объекты.

5 Технологическая схема

Технологическая схема подготовки включает четыре этапа (см. рис. 4). Первый этап имеет два входных потока – электронный и бумажный. Электронный входной поток создается авторами изобретений, которые используют программы формирования кодированных электронных форм патентных заявок и их подачи в соответствующее патентное ведомство с использованием Интернета в защищенном режиме передачи цифровых объектов.

В настоящее время электронные формы патентных заявок перед передачей преобразуются, как правило, в **факсимильную форму** и/или формат PDF. Поэтому технология их обработки практически полностью совпадает с технологией обработки бумажных заявок – исключается только сканирование.

Бумажный входной поток включает патентные заявки, которые создаются в бумажной форме и отправляются в патентное ведомство обычной почтой. Там они сканируются, в них разделяются вербальные и невербальные информационные объекты, после этого тексты распознаются.

На выходе первого этапа формируются три основных потока:

- изображения формул, которые передаются на второй этап технологической схемы;
- поток изображений и их подписей, который передается на третий этап;
- текстовый поток, который передается на четвертый этап технологической схемы.

Основная задача второго этапа заключается в том, чтобы расширить пространство поиска за счет формульной информации. Для патентных электронных библиотек необходимо обеспечить поиск, в первую очередь, по структурным химическим формулам. Для этого на втором этапе часть формул обрабатывается по специальным методикам. На выходе второго этапа формируется кодированный поток формульных объектов, которые передаются на четвертый этап.

Основная задача третьего этапа состоит в расширении пространства поиска за счет надписей и подписей к изображениям. Основная задача четвертого этапа заключается в объединении потоков текстов, изображений и формул, а также в обработке связей между объединяемыми объектами. После объединения компонентов электронные документы поступают в электронную патентную библиотеку. После индексирования они становятся доступны для поиска.

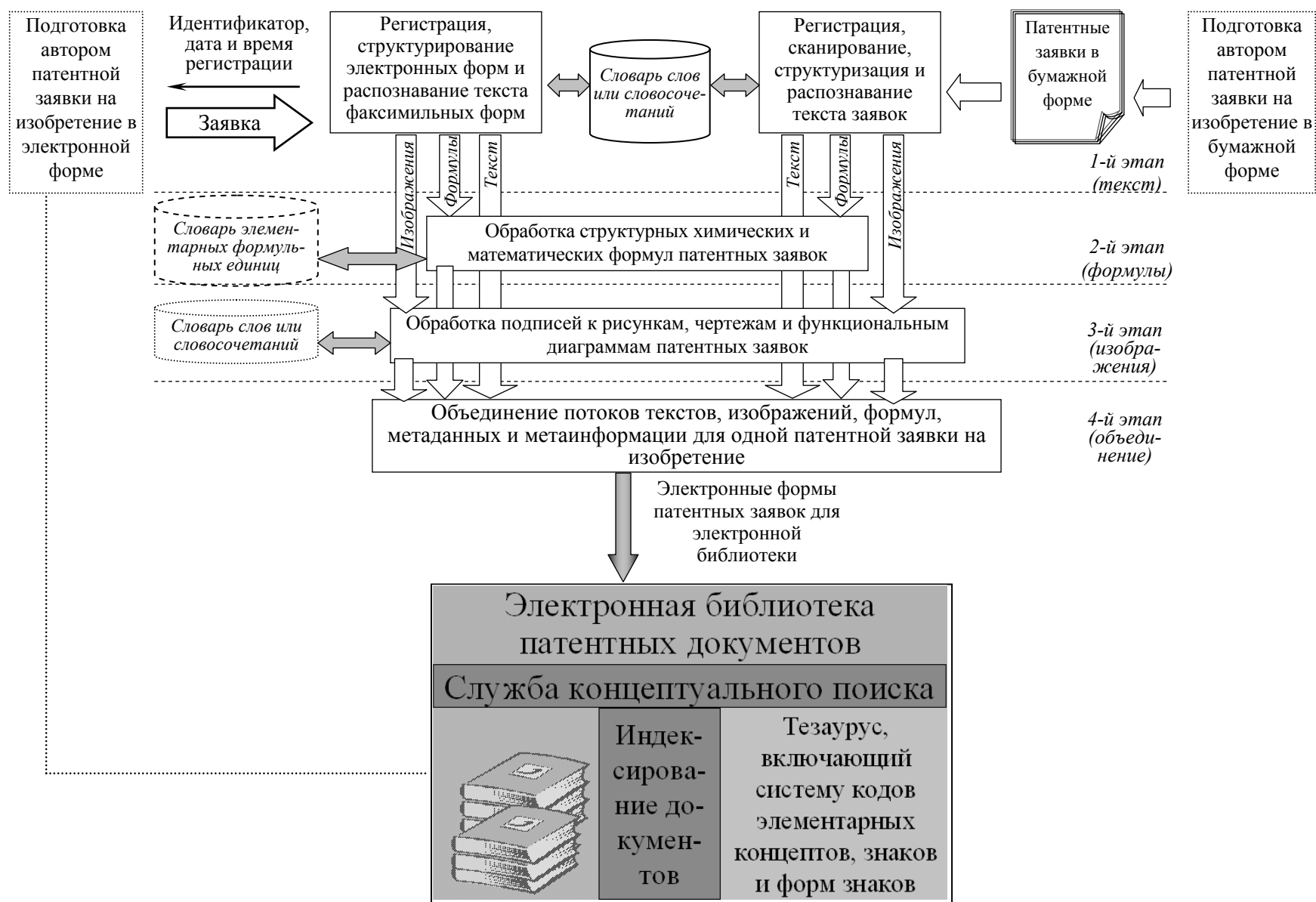


Рис. 4. Технологическая схема подготовки электронных форм патентных заявок

В докладе предполагается, что в процессах индексирования и поиска используется тезаурус, но при кодировании текста, формул и изображений он, как правило, не используется.

Индексирование электронных форм документов осуществляется после кодирования и объединения их компонентов. При выполнении операции индексирования доступны только цифровые **коды форм** вербальных и невербальных знаков, их сочетаний и/или цифровые **адреса** знаковых примитивов, например, **литер**.

Но **коды значений знаков** (в авторском понимании) и кодовые выражения для авторских концептов, которые эти формы выражают, в процессе индексирования сформировать с привлечением носителей авторских концептов невозможно, так как они (авторы) на этапе индексирования отсутствуют.

6 Развитие схем концептуального поиска

Предлагаемое развитие схем концептуального поиска (см. рис. 5) основано на выполнении операции кодирования концептов с использованием тезауруса в процессе подготовки автором электронных форм документов. Предлагаемое развитие не зависит от вида

документов – научные или патентные. Рис. 5 иллюстрирует случай комплексного кодирования научных документов.

Технологическая схема на рис. 4 показывает теоретическую возможность использования тезауруса при выполнении операций кодирования концептов. На этом рисунке возможность использования тезауруса на этапе подготовки автором электронных форм документов условно обозначена прямым углом из двух точечных линий.

Реализация комплексного кодирования имеет очевидные ограничения. Однако в тех предметных областях, в которых технология подготовки электронных документов авторами дает им возможность одновременно использовать тезаурус соответствующих электронных библиотек, предлагаемый подход технологически реализуем.

На рис. 5 различаются три основных вида концептов:

- искомые пользователем концепты (обозначены как U-концепты),
- интерпретированные пользователем концепты (I-концепты),
- авторские концепты (A-концепты).

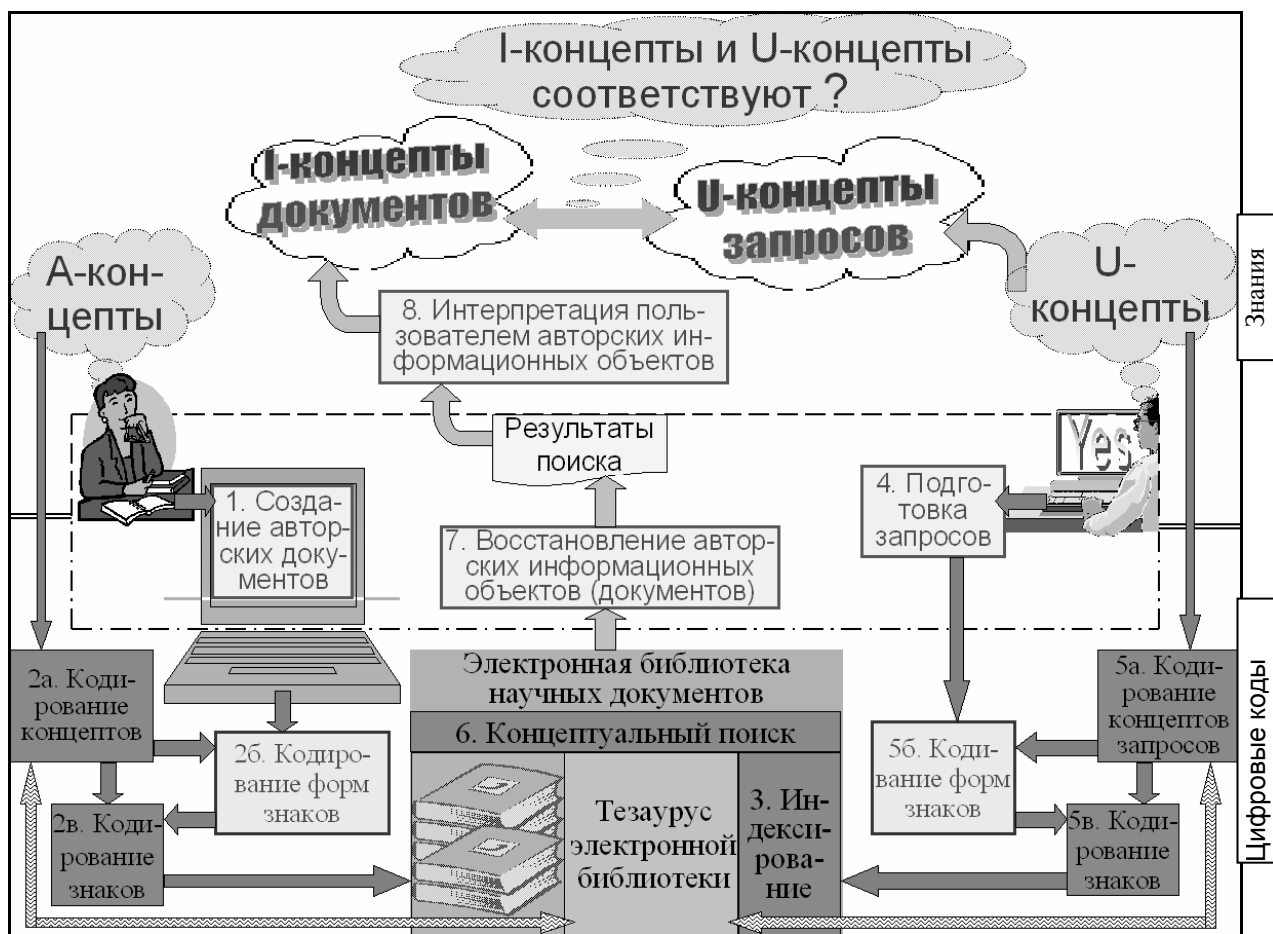


Рис. 5. Схема концептуального поиска с процессами комплексного кодирования

Комплексное кодирование документов, выполняемое автором до их отправки и формирования электронных библиотек, предоставляет принципиально новые возможности.

Во-первых, теоретически появляется возможность включить в электронные формы документов одновременно все три категории кодов:

- коды концептов, которые формируются в результате процесса семантического кодирования (см. процесс 2а на рис. 5),
- коды информационных объектов (форм вербальных и невербальных знаков), которые формируются в результате процесса информационного кодирования (см. процесс 2б на рис. 5),
- коды знаков, которые формируются в процессе знакового кодирования (см. процесс 2в на рис. 5).

Во-вторых, комплексное кодирование документов позволяет ограничить действие закона асимметрии средним уровнем, так как дескрипторы тезауруса выбирает автор и они, в общем случае, не зависят от форм представления содержания документа.

В-третьих, при выполнении процесса 2в появляется возможность кодирования тех знаков, формы которых не представимы в виде последовательностей дискретных примитивов, например, в виде последовательностей литер алфавитных систем письма.

Это возможность рассматриваемой схемы является принципиальным отличием от модели потоков (Stream Model) в 5S-подходе [3], в котором потоки (Streams) определены только как последовательности элементов произвольных типов.

На рисунке комплексное кодирование документов, выполняемое автором, разделено на три отдельных операции 2а, 2б и 2в. Кодирование запросов, выполняемое пользователем, также разделено на три операции 5а, 5б и 5в.

Предлагаемое развитие схемы не является решением задач соответствия концептов, а только одним из возможных подходов к их решению. Суть предлагаемого подхода заключается в замене информационного кодирования документов комплексным.

7 Пример комплексного кодирования

Рассмотрим пример комплексного кодирования изображений [6], использующий классификацию устьевых областей рек и их частей по морфологическим признакам из работы [7] и построим соответствующую ей знаковую систему. Один из возможных вариантов построения представляет собой сочетание из шести классов устьевых областей рек, перечисленных в третьем столбце табл. 1.

Будем использовать эту таблицу как определение знаковой системы устьевых областей рек, а также для построения соответствующих дескрипторов вербально-образного тезауруса электронной библиотеки, применяемого в процессе комплексного кодирования.

Перечисленные в третьем столбце таблицы названия устьевых областей будем использовать в качестве названий знаков. В работе [7] они сопровождаются стилизованными изображениями – пиктограммами, которые будем использовать в качестве форм образных знаков и образных дескрипторов тезауруса, обозначающих классы устьевых областей рек.

Зададим следующие правила назначения кодов каждому знаку знаковой системы устьевых областей рек, его форме и значению (концепту).

Их значениям присвоим коды 001к–110к (см. табл. 2) в соответствии с порядком перечисления названий знаков в третьем столбце табл. 1. Значения, объединенные с образными формами в виде пиктограмм (см. рис. 6), предлагается использовать как образные дескрипторы тезауруса. 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.

Шесть значений с кодами 001к–110к, которые объединены с названиями на русском языке, предлагается использовать как вербальные дескрипторы русскоязычной части тезауруса. Обозначим их коды как 001др–110др. Родовое понятие для них выражается дескриптором «устьевая область реки».

Устьевой участок реки	Устьевое взорье	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 [7]

Коды концептов устьевых областей рек	Названия устьевых областей рек на русском и английском языках и коды названий (в скобках)	Коды пиктограмм (форм образных дескрипторов, приведенных на рис. 6)
001к	Простая без блокирующей косы (001р) – Simple river mouth without blocked spit (001a)	001о
010к	Эстуарная без блокирующей косы (010р) – Estuary without blocked spit (010a)	010о
011к	Эстуарная с блокирующей косой (011р) – Estuary with blocked spit (011a)	011о
100к	Эстуарно-дельтовая без блокирующей косы (100р) – Silt delta without blocked spit (100a)	100о
101к	Эстуарно-дельтовая с блокирующей косой (101р) – Silt delta with blocked spit (101a)	101о
110к	Дельтовая с дельтой выдвижения (110р) – Protruding delta (110a)	110о

Табл. 2. Коды концептов и трех форм их представления: две вербальные и образная формы

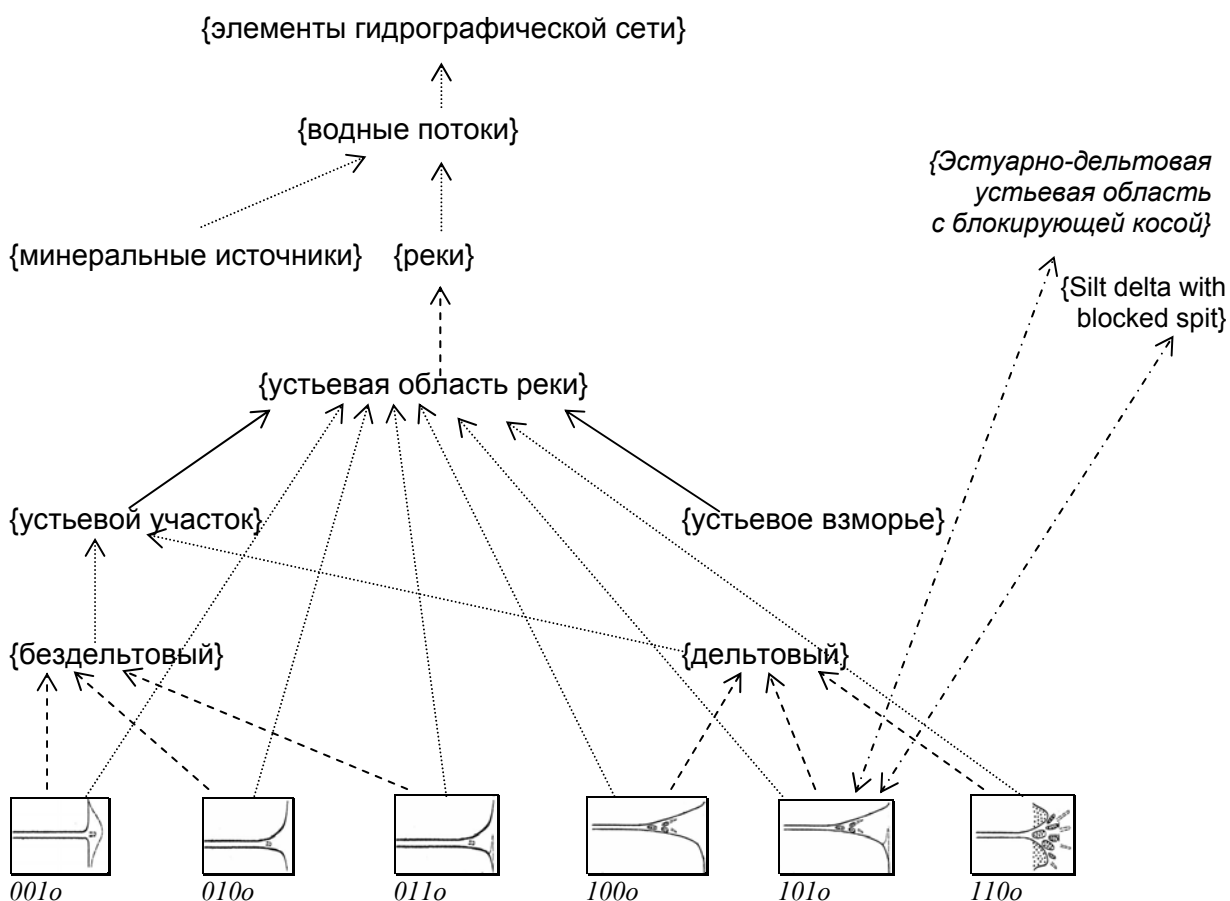


Рис. 6. Фрагмент вербально-образного тезауруса

Шесть значений с кодами 001к–110к, которые объединены с названиями на английском языке, предлагается использовать как вербальные дескрипторы англоязычной части тезауруса. Обозначим их коды как 001да–110да. Родовое понятие для них выражается дескриптором «river mouth» (на рис. 6 этот дескриптор не показан).

Отношения между дескрипторами наследуют отношения между этими геообъектами в предметной классификации. Фрагмент вербально-образного тезауруса на рис. 6 включает формы дескрипторов и отношения между ними.

Точечными стрелками на рисунке условно обозначены родовидовые отношения, сплошными – отношения «часть–целое», штриховыми –

ассоциативные отношения, штрих-пунктирными – отношения мультимодальной синонимии между дескрипторами разных информационных модальностей (на рис. 6 показаны отношения мультимодальной синонимии между двумя вербальными и одним образным дескрипторами).

8 Заключение

В технологической схеме подготовки электронных форм патентных заявок, существуют два входных потока: электронный и бумажный. Комплексное кодирование описаний заявок на этапе их подготовки автором возможно только при наличии их электронных форм и он-лайн доступе авторов заявок к тезаурусу электронной патентной библиотеки.

Возникает естественный вопрос: «Зачем автору, кроме подготовки заявок, дополнительно кодировать их содержание с помощью дескрипторов тезауруса?» Прежде, чем ответить на этот вопрос, обратимся к опыту промышленно используемой системы электронной подачи заявок, применяемой в Европейском патентном ведомстве (ЕПВ). С 2002 года ЕПВ разрешило авторам по их желанию добавлять в пакет электронных факсимильных документов их версии в оригинальных форматах редакторов Word или WordPerfect [8]. Основная причина, почему автор сам дополнительно включает версии документов в оригинальных форматах, заключается в том, что это позволяет ему в будущем предъявить эти версии в качестве доказательства программных ошибок формирования факсимильных документов, если такие будут обнаружены. В случае обнаружения ошибок и при наличии в пакете исходных документов в оригинальных форматах автор имеет право направить в ЕПВ запрос на исправление ошибок в заявках без изменения даты приоритета своей заявки.

Однако в случае обнаружения фрагментов патентных заявок, допускающих многозначное толкование, формализованная процедура выбора того значения, которое имел в виду автор в момент подачи заявки, в настоящее время отсутствует.

Комплексное кодирование описаний заявок на этапе их подготовки автором может служить основой для разработки формализованной процедуры фиксации нужного автору значения, включающей использование дескрипторов тезауруса и отношений между ними. Однако при подобном использовании тезауруса в процессе подготовки заявок необходимо учитывать динамику его изменения и способы фиксации динамики.

Одним из возможных способов учета динамики может быть явное указание даты и времени включения каждого нового дескриптора и установления каждой новой тезаурусной связи. Кроме указания даты и времени, необходимо с помощью специального атрибута помечать те новые дескрипторы и связи, появление которых в тезаурусе фиксирует появление в электронной библиотеке документов с новыми концептами, что в

существующих патентных электронных библиотеках не фиксируется. Таким образом, комплексное кодирование дает возможность автору зафиксировать с помощью дескрипторов тезауруса нужные ему концепты для фрагментов заявки, допускающих многозначное толкование.

Литература

- [1] Зацман И.М. Концептуальный поиск и качество информации.– М.: Наука, 2003.
- [2] Eco U. A Theory of Semiotics.– Bloomington: Indiana University Press, 1976.
- [3] Fox E.A. Digital Libraries of the Future: Integration through the 5S Framework // Труды 6-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции».- Пушкино: Ин-т математических проблем биологии РАН, 2004.- С. 1-2.
- [4] Шемакин Ю.И. Тезаурус в автоматизированных системах управления и обработки информации.- М.: Воениздат, 1974.
- [5] Гак В.Г. Асимметрия // Большой энциклопедический словарь «Языкознание».– М.: Большая российская энциклопедия, 1998.– С. 47.
- [6] Zatsman I. Pictorial Signs for Geomages in Digital Libraries // European Journal for Semiotic Studies. Vol. 15. N. 2–4. 2003. P. 609–620.
- [7] Михайлов В.Н. Гидрология устьев рек: Методическое пособие. М.: Изд-во МГУ, 1996.
- [8] Notice dated 1 June 2002 concerning the electronic filing of European patent applications and subsequent documents // Official Journal of the EPO. 2002. N 7.- P. 372–373.

Semantic, information and sign coding of patent documents in digital libraries

Igor M. Zatsman

The paper is devoted to the outline of conceptual search in digital libraries of patent documents. This study is based on the problem setting of conceptual search in digital libraries, suggested in the monograph [1]. It contains the description of two variants of the outline of conceptual search. Variants differ in sources of information resources that are used for formation of digital libraries, namely - the person and/or technical system [2]. The objective of the paper is to define a procedure of the multiobject coding of documents. The multiobject coding consists of

- the semantic coding of document content;
- the information coding, in other words, the coding of sign forms in documents;
- the sign coding.

The sign is a unity of its form and content here.