Организация и использование информационных ресурсов

АНТОПОЛЬСКИЙ Александр Борисович *– доктор технических наук, директор Некоммерческого партнерства «Электронные библиотеки»*

АУССЕМ Владимир Игоревич *– главный специалист НТЦ «Информрегистр»*

ТИПОЛОГИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ В СТАНДАРТНЫХ СИСТЕМАХ МЕТАДАННЫХ: АНАЛИЗ И ПРОБЛЕМЫ ИНТЕГРАЦИИ [1]

Среди большого количества разнообразных систем метаданных, используемых в информационных ресурсах в сфере науки, культуры и образования и рассмотренных нами в предыдущих публикациях [1,2], наиболее распространенными и признанными являются системы метаданных семейства MARC, Дублинское ядро метаданных, а также быстро распространяющаяся в информационно-образовательной среде система метаданных LOM. Поэтому при создании федеративных электронных библиотек, а также сводных каталогов электронных ресурсов их разработчикам чаще всего приходится иметь дело с интеграцией описаний ресурсов представленных в этих системах метаданных. Это влечет за собой необходимость сопоставления указанных систем и поиска способов интеграции описаний ресурсов, представленных в указанных стандартах. Действительно, как в России, так и за рубежом разработан ряд моделей отображения указанных систем между собой [3,4.]. Однако эти отображения носят в основном формально-синтаксический характер, игнорируя особенности семантики представления различных характеристик ресурсов в этих стандартах систем метаданных. Такой же подход характерен для универсальной системы интеграции метаданных, развиваемой в рамках концепции семантического web на основе RDF-схем. Это объяснимо из-за сложности, порой необозримости задачи, но не решает главной задачи интеграции метаданных: обеспечения единой системы навигации по метаданным в федеративных электронных библиотеках.

Настоящее исследование имеет целью проанализировать подходы в перечисленных стандартах метаданных применительно к одной из важнейших характеристик (точнее группы характеристик), которую принято называть *типом* или *видом* информационных ресурсов. Тип ресурса в явном или неявном виде присутствует во всех коллекциях, содержащих неоднородные ресурсы или их описания, а также в большинстве развитых поисковых систем. В тоже время трактовка этой важной характеристики информационных ресурсов в различных системах метаданных также существенно различается.

Для сравнения приведем определение понятие «Тип ресурса» из официальных спецификаций двух стандартов метаданных

Спецификация RUS LOM

Тип ИР – комплексная характеристика ИР, отражающая его структуру, форму и способ реализации, вид и характер содержащейся в нем информации, а также иные свойства ИР, обусловливающие возможности его использования, но не учитываемые другими классификациями.

Спецификация DC

Элемент данных: Тип ресурса.

Определение: Природа или жанр содержания ресурса.

Комментарий: *Тип* включает термины, описывающие общие категории, функции, жанры или объединенные уровни содержания. Для практического использования рекомендуется выбирать значение из контролируемого словаря (например, рабочего проекта списка Типов Dublin Core DCT1). Для описания физического или цифрового представления ресурса используется элемент *Формат.*

Из приведенных определений с очевидностью следует, что понятие «тип ресурса», понимается, мягко говоря, неоднозначно. С еще большей очевидностью это следует при анализе словарей, предлагаемых для данного элемента метаданных, что будет сделано ниже.

Метаданные и классификации информационных ресурсов

В принципе, характеристики любого типа могут послужить основой для создания той или иной типологии информационных ресурсов. При этом наибольший интерес для анализа представляют классификационные характеристики систем метаданных, претендующих на универсальность, хотя бы в пределах крупной области деятельности (наука и инновации, промышленность, образование, издательская деятельность и т.д.). Именно они, по идее, должны отражать установившиеся общие представления о структуре и составе информационной сферы, в то время как часто модифицируемые классификационные деления многочисленных каталогов интернет-сайтов, поисковых машин, периодических и продолжающихся изданий строятся по принципу ad hoc для конкретного информационного массива или меняющегося новостного потока.

3. Уровни и аспекты описания информационных ресурсов

Информационные ресурсы, как правило, являются сложными объектами с многоуровневой структурой. Обычно можно выделить, например, такие уровни, как уровень ИР в целом, самостоятельные

[1] Настоящее исследование поддержано грантом РФФ № -07-90087

части ИР (например, разделы сайта или отдельные документы), тексты (рисунки, фотографии и т.д.), предложения, слова и числа, знаки (идеограммы, точки изображения и т.д.). На каждом из таких уровней мы в принципе имеем дело с разными объектами, имеющими собственные характеристики, охватываемыми разными, сложившимися в различных научных дисциплинах классификациями. Тем не менее, при описании информационного ресурса часто используются характеристики "низших" уровней.

Возможные аспекты рассмотрения, характеристики объектов разного уровня могут и должны значительно различаться. Так, на уровне знаков практически[2] бессмысленно говорить о тематике, но уже на уровне частей документа (и, возможно, на уровне лексического состава) это вполне целесообразно. В то же время для больших библиотечных коллекций, политематических баз и банков данных тематическая характеристика снова теряет смысл.

К сожалению, основные системы метаданных крайне редко четко выделяют и называют аспекты рассмотрения информационных ресурсов, классификационные основания выделения типологических (классификационных) характеристик, а также уровни описания. Исключение представляют, пожалуй, лишь ГОСТы 7.60-90 и 7.83-2001 [5,6], однако они классифицируют лишь издания (традиционные и электронные). В результате, метаданные, относящиеся к разным аспектам рассмотрения, разным классификационным основаниям, разным объектам, перечисляются авторами в едином ряду, что иногда затрудняет понимание и использование системы метаданных в целом. Нередко метаданные, выделенные по разным аспектам и основаниям, получают в разных системах метаданных одинаковые наименования. Наиболее велика степень омонимичности употребления названия "*тип (вид) ИР*". Есть, однако, и более тонкие и, следовательно, более сложные случаи. Так, в DC метаданное "Text" охватывает факсимиле и изображения текста (например, в формате pdf), а в системах MIME и UNIMARC – исключает.

На основе анализа распространенных систем метаданных (в том числе упомянутого ГОСТ) можно предложить следующий примерный перечень аспектов рассмотрения информационных ресурсов и оснований для их классификаций:

физическая природа (материальная конструкция),
знаковая природа,
целевое назначение,
жанр (виды документов и т.п.).

Типология физической природы информационных ресурсов

В данном случае речь идет фактически о классификации носителей информации, которые, вообще говоря, могут быть необъятно разнообразными. История знает и наскальные изображения, и сложные технические объекты, созданные для манифестирования определенной информации (например, маяки). Нас, однако, будут интересовать в основном современные машиночитаемые носители информации[3]. В определенном смысле можно считать, что указание носителя характеризует информационную систему, в которой используется информационный ресурс, а не сам ресурс. Так, дискуссионным является вопрос о том, одним или разными ресурсами являются одинаковые по содержанию произведения (коллекции), записанные на разных носителях.

В принципе, задача классификации так называемых технических носителей представляется менее сложной, чем, в ряде случаев, носителей "традиционных". Идентификация первых, являющихся продуктами современного промышленного производства, опирается (даже для малосерийных изделий) на развитую систему марок, технических наименований, спецификаций.

Некоторые распространенные системы метаданных объединяют данные о носителе с другими техническими и технологическими характеристиками. Таковы, например, в системе LOM классификаторы типов и названий "технологий и программных сред, требуемых для использования компьютерного ИР". Они в основном лишь опосредованно определяют физическую форму ресурса.

В DC носители могут быть указаны в элементе "Format", определяемом как "физическая или цифровая форма ресурса". При этом данный элемент "может быть использован для идентификации программного обеспечения, компьютерного или иного оборудования, необходимого для размещения или эксплуатации ресурса". Документы DC рекомендуют использовать для данного элемента контролируемые перечни значений, но не указывают на конкретный классификатор носителей информации, в отличие, например, от перечня форматов записи. Можно считать, что это дает возможность использовать "чужие" наименования видов носителей при описании в терминах DC ресурсов, ранее описанных другими системами метаданных.

---

[2] Наличие в тексте отдельных "характерных" знаков (химических, математических и т.д.) может почти однозначно свидетельствовать о его общей тематической направленности, но речь здесь может идти, скорее, о "маркере" тематики, а не о ее точном описании.

[3] Несмотря на быстрый рост разнообразия их видов, классификация машиночитаемых носителей еще не достигла сложности и разветвленности классификации по рассматриваемому основанию традиционных изданий (см. раздел "Виды изданий по материальной конструкции" ГОСТ 7.60.90), особенно, если к ним добавить кино-, фото и аудиодокументы (см. ГОСТ 7.72-96 [7]).

DC, тем не менее, имеет специальное обозначение для одной, весьма специфической физической формы информационного ресурса. Речь идет о значении "Физический объект" элемента "Тип"[4]. Однако, судя по определению, эта форма не имеет отношения к электронным информационным ресурсам.

Другие системы метаданных задают ограниченный перечень классов носителей информации. Таковы, например, позиция "Специальное обозначение материала" ($a/1) в поле кодированных данных для электронных ресурсов (поле 135) в UNIMARC и элементы книготорговой системы метаданных ONIX "код и описание формы продукции", "код и описание типа электронной книги" из раздела "Область идентификатора издания". Приведем перечень обозначений материала в UNIMARC:

кассета с магнитной лентой;
компьютерная микросхема в кассете;
кассета с компьютерным оптическим диском;
кассета с компьютерной магнитной лентой;
магнитная лента для мэйнфреймов;
компьютерный гибкий диск;
компьютерный магнитооптический диск;
компьютерный оптический диск;
онлайновые системы.

В связи с быстро развивающейся технологической ситуацией жестко заданные перечни значений для рассматриваемого элемента описания ИР должны оперативно актуализироваться, что происходит, как видно, в том числе и из приведенного примера, далеко не всегда. В целом же представляется целесообразным ориентироваться на официальные стандартизованные перечни и классификации типов носителей, например, на ГОСТ 7.72-96, при условии его своевременной актуализации.

Типологии семиотической природы информационных ресурсов

Семиотические характеристики, строго говоря, обычно относятся не к ресурсу в целом, а к отдельным знакам, с помощью которых отражается содержание входящих в ресурс произведений. Таким образом, характеристика семиотической природы информационного ресурса может формироваться только как общая, преобладающая или обобщенная характеристика содержащихся в нем отдельных знаков.

Самая сложная проблема при определении семиотических характеристик электронного информационного ресурса являются множественность его представлений в компьютерной системе и многоуровневость знаковой структуры. Анализируя эту структуру, можно в пределе дойти до рассмотрения двоичного кода. Представляется более разумным для характеристики ресурса ориентироваться на его представление, воспринимаемое человеком. С другой стороны, не менее обоснована ориентация на программно-операционный уровень, то есть на набор знаков, разные типы которых по-разному воспринимаются и обрабатываются используемым системным программным обеспечением.

Иллюстрацией может служить уже приводившийся пример с интерпретацией изображения текста в формате pdf – в DC это текст, а в UNIMARC и MIME – изображение. Таким образом, DC "смотрит" на систему знаков "по-человечески", а UNIMARC и MIME – "по-программному".

Распространенные системы метаданных обычно дают возможность обозначения следующих основных классов знаков:

Текст (то есть буквенные записи текстов на естественных языках и другие, встречающиеся в "обычных" публикациях символы),

Изображение (обычно есть возможность различить неподвижное и подвижное изображения),

Звук.

Иногда выделяют еще *цифровые записи* (например, в UNIMARC[5]). Более глубокие семиотические классификации, о возможности построения которых упоминалось в начале раздела, обычно не используются[6].

Среди значений элемента DC "Тип" есть еще два, в определенной степени характеризующих семиотическую природу информационного ресурса. Это термины "Data set" и уже упоминавшийся "Physical Object". "Data set" означает, по определению, наличие определенной знаковой структуры, то есть семантической характеристики, связанной с прямой зависимостью смысла знаков или их сочетаний от места в этой структуре. Что касается физических объектов, то нельзя, например, отрицать, что скульптура в целом или отдельные ее фрагменты должны рассматриваться как знаки, несущие определенную смысловую и

[4] Приведем полное (вместе с официальным примечанием) определение этого значения (термина): "Трехмерный объект или субстанция, – например, компьютер, великая пирамида, скульптура. Заметим, что его цифровое представление или иные его суррогатные формы должны описываться с использованием терминов "Изображение", "Текст" или других значений элемента "Тип".

[5] Здесь UNIMARC, однако, придерживается "человеческой" точки зрения: "цифровым" называется "файл, который *преимущественно* содержит цифровые данные …" (курсив мой).

[6] Весьма широкий (но не глубокий иерархически) перечень типов MIME определяет, все-таки, различия в способах обработки записей разными программными средствами, а не в способах передачи их смыслового или эстетического содержания.

эстетическую информацию. Поэтому можно считать, что во многих случаях (но не всегда) термин DC "Physical object" характеризует не только физическую, но и семиотическую природу информационного ресурса.

Системы метаданных могут использовать для отражения семиотических характеристик разные разделы (группы метаданных), иногда даже в рамках одной модели. Так, помимо отражающих семиотическую структуру значений элемента "Тип", в DC имеется упоминавшийся ранее специальный элемент "Format", в записи которого среди прочего можно отразить и воспринимаемый и обрабатываемый программным обеспечением формат записи[7], то есть семиотическую характеристику (правда, на этот раз уже "с точки зрения программы"). Заметим при этом, что среди значений элемента "Тип" DC имеются и не относящиеся напрямую к семиотическим характеристикам.

В UNIMARC основные семиотические типы для электронных информационных ресурсов[8] (*текст, изображение, цифровая запись*) символизируются кодами позиции "вид электронного ресурса" ($a/0). Здесь же возможно указание на "комбинированные" и "другие" файлы. В то же время наличие цвета и звука отмечается в других позициях ($a/2 и $a/4, соответственно) поля кодированных данных "*электронные ресурсы*" (поле 135). В том же поле есть позиции для технических характеристик, в значительной мере отражающих возможности и границы семиотической системы ресурса и ее типологическое разнообразие. Это характеристики *"битовая глубина изображения", "количество файловых форматов" и "уровень сжатия".*

Нередко семиотическая характеристика информационного ресурса становится ясной из обозначения жанра входящих в него произведений (традиционная характеристика "*вид документа*"). Так, если в описании ресурса в модели LOM он характеризуется как "narrative text", то, очевидно, что "Тип" DC должен быть определен как "Text". "Graph", соответственно, определяется как "Image", а "table" как "Dataset".

Таким образом, обозначения основных семиотических характеристик в распространенных системах метаданных либо прямо совпадают (*текст, изображение, звук*), правда, иногда с точностью до "точки зрения", либо (в значительной части оставшихся случаев) выводятся из значений других характеристик.

Классификации информационных ресурсов по жанру

Жанровые характеристики являются, пожалуй, самыми разнообразными и синтетическими параметрами информационных ресурсов. Собственно говоря, они относятся не к ресурсу в целом, а к отдельным произведениям (документам), входящим в его состав. Однако так как подавляющее большинство описываемых в распространенных каталогах ресурсов являются моновидовыми (на определенном уровне типизации), жанр (для традиционных документальных систем чаще используется термин "вид документа") служит основой некоторых классификаций информационных ресурсов и отражается многими системами метаданных.

Указание жанра может одновременно характеризовать особенности содержания ресурса (произведения), назначение, структуру текста и, как уже отмечалось, знаковую природу ресурса. Традиционно, в разных сферах деятельности складывались собственные типологии документов, в разной степени стандартизованных. Напомним хотя бы о видовых классификациях научно-технической информации (на которых основана система стандартов СИБИД), видов печатных изданий (ГОСТ 7.60.90), Единой системе конструкторской документации (ЕСКД), различных типологиях официальных и деловых документов (в том числе в системах документооборота и т.д..

С развитием крупных информационных систем, охватывающих разнородные информационные объекты, и появлением обобщающего понятия "информационные ресурсы" информатике не только пришлось окончательно освоить жанровые классификации фотографии, кино, радио и телевидения, но и обратиться к таким экзотическим для документалистики областям, как изобразительное искусство, электронные модели и тренажеры, компьютерные игры. При этом каждая из этих областей имеет сложившуюся, если не в строгом научном обороте, то в разговорно-журналистской и рекламной практике[9], жанровую (видовую) типологию.

В результате, объединенный перечень наименований жанров для информационных ресурсов становится практически необозримым. Поэтому неудивительно, что универсальные системы метаданных не всегда пытаются задать контролируемый словарь видов (форм, жанров) документов (произведений) для современных электронных ресурсов.

Так, в UNIMARC имеется ненормированное поле "примечание о виде электронного ресурса" (поле 336), где приводится "специфическая информация, такая как форма и жанр текстового материала". Примеры: биография, словарь, указатель, спортивная статистика, дайджесты. Правда, в "основной части" всех моделей семейства MARC используются развитые традиционные классификации видов изданий.

В DC практически нет специальных средств для отражения жанровых особенностей материала. Остается, конечно, возможность отразить их в свободной записи элемента "Description", которая чаще всего

[7] При этом прямо рекомендуется (но не жестко предписывается) использовать для этого перечень типов файлов системы MIME.

[8] Для "традиционных" изданий в MARC используется более развитая семиотическая классификация.

[9] Вспомним, например, о "бродилках", "стрелялках" и т.п. для компьютерных игр

представляет собой обычный реферат или аннотацию. Однако соответствующая рекомендация (а, тем более, указание) отсутствует. Тем не менее, среди значений элемента "Тип" DC имеется, по крайней мере, одно, явно относящееся к жанровой характеристике содержания ресурса, хотя и на весьма высоком уровне обобщения. Это – термин "Event"[10]. Примеры "событий": *выставка, конференция, совещание, "день открытых дверей", представление, сражение, суд, свадьба* и т.д.

Система метаданных LOM, предназначенная для описания хотя и весьма широкой, но все-таки в определенной степени ограниченной области образовательных ресурсов, рискует задать ограниченный "жанровый" перечень в словаре для элемента "Тип образовательного ИР"[11]: *упражнение; модель, среда моделирования, тренажер; массив вопросов; схема, чертеж, диаграмма; рисунок, иллюстрация; график; указатель, оглавление; кадр, слайд; таблица; описательный текст; экзамен, тест; средства выполнения учебно-исследовательского эксперимента; формулировка проблемы (задачи);*

материал для самоконтроля знаний или умений; лекция (конспект).

В целом из-за обширности жанровых классификаций представляется разумным не создавать самодеятельные волей-неволей усеченные и субъективные перечни, а либо допустить использование свободных записей, либо ориентироваться на устоявшиеся в конкретных областях деятельности стандартизованные перечни видов документов и иных материалов. В последнем случае могут возникать сложности, во-первых, связанные с разной степенью и разной глубиной традиции "официальной" стандартизации в разных областях, прежде всего, для объектов относительно новых в рамках понятия "информационные ресурсы" (например, модели, сервисные и интерактивные системы и т.д.). Во-вторых, в универсальных и достаточно широких по охвату системах метаданных придется указывать описываемую область деятельности и применения ресурса и присущий ей стандарт видов документов (информационных объектов).

Назначение ресурса

Указанием на область применения ресурса может служить характеристика целевого назначения, которая является основанием для одной из традиционных классификаций видов изданий (раздел "Виды изданий по целевому назначению" (ГОСТ 7.60.90, а также ГОСТ 7.83-2001).

Большинство распространенных в настоящее время систем метаданных не выделяют в явном виде специальных элементов для обозначения назначения ресурса. Одним из немногих исключений является система метаданных Государственного регистра баз и банков данных (НТЦ "Информрегистр"). Объяснением, вероятно, может служить тот факт, что обычно классификация по назначению соответствует (а то и словесно совпадает) до некоторого уровня с видовыми классификациями. В самом деле, например, термин "чертежно-конструкторский" может быть применен и в "целевой", и в "видовой" типологиях. В приведенном ранее перечне "типов" LOM многие термины явно соответствуют выражаемым словесно несколько иначе указаниям на уточнение цели использования ресурса, в общем виде (образовательные информационные ресурсы) уже сформулированной в наименовании модели.

Есть, однако, один класс ресурсов, выделяемый фактически по признаку назначения и отражаемый почти во всех системах описания электронных ресурсов. Это – компьютерные программы, software. Так, в DC это одно из значений элемента "Тип", а в UNIMARC – позиции "вид электронного ресурса". Также характеристикой назначения можно считать термин "Service" из словаря элемента "Тип" DC. Это тем более имеет смысл из-за быстрого развития и постепенной стандартизации системы видов "служб" и "электронных услуг".

Заключение

Таким образом, сравнительный анализ характеристик ресурсов, связанных с понятиями *тип* и *вид* в стандартных системах метаданных показывает, что иерархическая таксономическая классификация этих понятий невозможна из-за разных оснований деления, используемых для этих понятий в разных системах метаданных. Выходом для интеграции этих понятий в федеративных электронных библиотеках является использование фасетной классификации с четко и эксплицитно выделенными основаниями деления (меронами). В настоящей статье сделана предварительная попытка разработки такого перечня меронов. В ближайшем будущем в рамках исследований, проводимых в Российской ассоциации электронных библиотек, будет предложена практическая методика соотнесения характеристик понятий предлагаемых в стандартах метаданных для *типа* и *вида* ИР с рекомендуемым перечнем меронов.

Литература

1. Антопольский А.Б., Ауссем В.И., Блау С. А., Жежель А. И. Отчет о результатах работ 2004 г. по проекту РФФИ 04-07-90087 “Исследование и разработка системы метаданных для электронных информационных ресурсов и сервисов в фундаментальной науке”. - www.inforeg.ru, 2005

[10] "Метаданные для термина "Event" предоставляют описательную информацию, на основе которой можно определить цель, место, продолжительность события, ответственных лиц и связи с другими событиями и ресурсами".

[11] Далее приводится русский перевод значений элемента 5.2 learning resourcetype.

2. Соколовский В.В. Сопоставительный анализ стандартов метаданных для сферы науки, культуры и образования / В.В. Соколовский. - М., 2006. - 12 с. (Серия "Электронные библиотеки: теория и методика"; Препринт № 06.004, март 2006 г. / Российская ассоциация электронных библиотек - НП ЭЛБИ).

3. Метаданные для информационных ресурсов сферы образования. Руководство по применению информационной модели и ее XML-привязки. Версия 1.0, Москва 2006 www.spec.edu.ru.

4. Dublin Core Metadata Element Set Mapping to Mods. Version 3.0? 2003 www.loc.gov/standards/mods/dcsim/ple-mods.html

5. ГОСТ 7.60-90. Система стандартов по информации, библиотечному и издательскому делу. Издания. Основные виды. Термины и определения.

6. ГОСТ 7.83-2001. Система стандартов по информации, библиотечному и издательскому делу. Электронные издания. Основные виды и выходные сведения.

7. ГОСТ 7.72-96. Система стандартов по информации, библиотечному и издательскому делу. Коды физической формы документов.