Метаданные, их свойства, функции, классификация и средства представления

М.Р. Когаловский
Институт проблем рынка РАН

XIV Всероссийская научная конференция RCDL-2012. Переславль-Залесский, 15-18 октября 2012 г.

Цель доклада

- *«Метаданные»* - один из наиболее популярных терминов в ИТ в настоящее время
- Употребление его особенно активизировалось с появлением Веб
- Но трактовка этого термина не устоялась до сих пор
- Метаданным посвящены тысячи публикаций, но большинство посвящено обсуждению конкретных стандартов
- Мало публикаций концептуального характера
- Существуют заблуждения, касающиеся свойств и функций метаданных, хронологии возникновения термина
- Основная цель доклада – обсудить смысл термина метаданные, свойства и функции информационных ресурсов этого вида
- Метаданные - особый вид информационных ресурсов
- Метаданные «горизонтальной» сферы («универсальные») и метаданные «вертикальной» сферы «специализированные»).

Немного истории - 1

- Метаданные начали использоваться в ИТ задолго до рождения термина
- Несколько примеров:
  - ✓ описания типов данных в программах на языках программирования
  - ✓ описания файлов, встроенные в программы и в спецификациях на языках управления заданиями (IBM JCL)
  - ✓ описания форматов отчетов в языке IBM RPG
  - ✓ поисковые образы документов в ранних ИПС
  - ✓ диаграммы потоков данных в CASE-инструментах и др.
- Данные в операторах языков программирования и др. языков: *description, definition, declaration (и не только в них)* – разновидности метаданных
- Когда возник термин *метаданные*?
- Одно из странных распространенных заблуждений: этот термин возник в 1999 г., когда директорат DCMI опубликовал спецификацию DC 1.1
- Можно было ожидать, что термин *метаданные* родился в области технологий баз данных.

Немного истории - 2

- В работе: *James Fry, Davis W. Jeris. Toward a Formulation and Definition of Data Reorganization /SIGMOD Workshop **1974**. Ann Arbor, Michigan* ошибочно утверждается, что термин был введен в статье: *G.H. Mealy. «Another Look at Data». Proc.**1967** FJCC, AFIPS vol. 31.*
- В статье *«Metadata»* англоязычной Википедии ссылаются на отчет: *P. Bagley. Extension of Programming Language Concepts. Philadelphia: University City Science Center, November **1968**.* В нем действительно используется термин *метаданные*.
- Статистический анализ *ACM SIGMOD Anthology* (с ретроспективой от 1969 г.) показал: в области БД термин *метаданные* начал активно использоваться на пороге 1980-х гг., хотя использовался и ранее
- Дальнейшая активизация его использования связана с рождением XML-технологий и концепции Semantic Web; в их контексте, он чаще всего трактуется как описание контента информационных ресурсов
- Сфера применения метаданных очень широкая, электронные библиотеки – лишь одна из областей, где они используются.

Об определении термина метаданные

- Что же такое *метаданные*? Проблемы лаконичности определения.
- Много различных трактовок в литературе, большинство не являются достаточно полными или даже ошибочны
- Ряд примеров определений - в тексте статьи *(см. Труды RCDL-2012)*
- Распространенное определение *«Метаданные – это данные о данных»* не охватывает все виды метаданных, используемых в современных ИТ, и в малой степени содержательно
- В работе *Ling Liu, M. Tamer Ozsu (eds.). Encyclopedia of Database Systems. Springer, 2009. 748 p. (*870 авторов, более 3000 статей) дано определение: *«Метаданные – это данные, связанные с каким-либо элементом данных»*. Понятие *элемент данных* не определяется , ряд положений статьи подвержен критике, сосуществуют *Metadata* и *Meta data*; не определены термины *Data*, *Database*, *Data Model*, но определяются составные термины
- В энциклопедии *A. Ralston, E.D. Reylly, D. Hemmendinger (eds.). Encyclopedia of Computer Science, 4th edition. John Wiley & Sons Ltd, 2003. - 2034 p.* термин *метаданные* не определяется. Не определяются и *Data, Information; Database* – определяются только реляционные.

Почему не устоялось определение термина

- Большое многообразие видов метаданных, обусловленное:
  - ✓ множеством сфер применения с различными потребностями пользователей
  - ✓ разнообразием природы описываемых ресурсов
  - ✓ разнообразием подходов к представлению метаданных
- Недостаточная компетентность авторов публикаций, посвященных метаданным, которая вводит в заблуждение читателей: во многих публикациях рассматривается частный вид метаданных, не делается должных оговорок, в результате неправомерно обобщаются их свойства, присущие частному случаю
- Большое количество сообществ занято созданием систем метаданных
- Существование двух, все еще не согласованных подходов к пониманию смысла метаданных:
  - ✓ подход библиотечного сообщества, истоки которого - в технологии документальных ИПС (это, главным образом, метаданные текстовых систем)
  - ✓ подход сообщества CS, истоки которого - в области технологий баз данных и др. направлений ИТ, связанных с управлением данными и знаниями.

Часто встречающиеся заблуждения

- Метаданные могут быть только у структурированных данных
- Метаданные являются структурированными даннымиуп
- Метаданные для Веб являются слабоструктурированными данными
- Функция метаданных - описание семантики информационных ресурсов
- Семантические метаданные могут использоваться только для структурированных данных
- Метаданные – это данные о данных
- Ошибочные представления о времени рождения этого термина. Например, упоминавшееся утверждение о том, что термин *метаданные* появился в 1999 г., когда директорат DCMI опубликовал DC 1.1.

  Но: *NCSA/OCLC Metadata Workshop* (март 1995), результатом которого стало учреждение инициативы Дублинского ядра и создание DCMI.

NCSA = National Center Supercomputing Application
OCLC = Online Computer Library Center

Расширение видов описываемых ресурсов

- Первоначально метаданные использовались для описания разнообразных информационных ресурсов
- Теперь, вместе с тем, они описывают и ресурсов других видов:
  - ✓ пользователей систем (их профили)
  - ✓ авторов представленных в ЭБ публикаций
  - ✓ организации – создатели и/или владельцы информационных ресурсов либо ИТ-сервисов (например, владельца веб-сервиса в реестре UDDI)
  - ✓ концептуальные схемы предметных областей
  - ✓ онтологии предметных областей
  - ✓ интерфейсы веб-сервисов
  - ✓ бизнес-процессы
  - ✓ потоки работ
  - ✓ объекты на географических картах (символами легенды)
  - ✓ различные аспекты создаваемых систем (UML-диаграммы в CASE-инструментах).
- Именно учитывая такое более широкое назначение термина *метаданные,* правомерно использовать более общий термин **метаинформация.**

Примеры метаданных - 1

- *В технологиях баз данных:*
  - ✓ концептуальные схемы предметных областей
  - ✓ схемы баз данных
  - ✓ описания междууровневых отображений схем в системах баз данных
- *В технологиях интеграции данных:*
  - ✓ локальные схемы источников данных
  - ✓ глобальные схемы
  - ✓ описания отображений между локальными схемами интегрируемых источников и глобальной схемой
  - ✓ онтологии локальных источников и общей онтологии системы интеграции данных
  - ✓ описания отображений между онтологиями локальных источников и общей онтологией
  - ✓ характеристики регистрации источников в посредниках в системах виртуальной интеграции данных.

Примеры метаданных - 2

- *В технологиях текстового поиска:*
  - ✓ идентификаторы текстовых документов
  - ✓ наборы значений индексирующих атрибутов документов
  - ✓ индексы коллекций документов в системах текстового поиска
  - ✓ библиографические описания документов
  - ✓ аннотации публикаций
  - ✓ каталоги коллекций документов
  - ✓ наборы ключевых слов документов
  - ✓ рубрики классификаторов для документов
  - ✓ наборы значений элементов метаданных DC
  - ✓ индексы УДК
  - ✓ индексы ISBN монографий
- *В CASE-технологиях:*
  - ✓ UML-диаграммы проектов разрабатываемых систем
  - ✓ диаграммы IDEF
  - ✓ ER-диаграммы.

Примеры метаданных - 3

- *В веб-технологиях:*
  - ✓ гипертекстовая разметка веб-страниц
  - ✓ наборы имен и значений параметров тегов META в веб-страницах
  - ✓ разметка фрагментов веб-страниц средствами микроформатов (hCard, hReview, hProduct, hRecipe и др.)
  - ✓ описания типов XML-документов (DTD)
  - ✓ XML-схемы для типов XML-документов
  - ✓ RDF-спецификации ресурсов
  - ✓ описания онтологий на языке OWL или OWL2
  - ✓ семантические аннотации веб-страниц или их фрагментов
- *В технологии веб-сервисов:*
  - ✓ описания интерфейсов веб-сервисов средствами языка WSDL
  - ✓ описание характеристик веб-сервисов в регистре UDDI
  - ✓ описание организаций-владельцев веб-сервисов в регистре UDDI.

UDDI = Universal Description Discovery & Integration (консорциум OASIS)
UDDI Registry – Microsoft, IBM и Ariba

Основные свойства метаданных - 1

- *Относительный характер* разделения информационных ресурсов на данные и метаданные
- *Разнообразие областей*, в которых используются метаданные, и видов описываемых ресурсов
- Зависимость свойств метаданных от характера использующей их системы, вида описываемых ресурсов, используемых ИТ, потребностей пользователей систем и т.п.
- Зависимость состава метаданных от *информационной архитектуры* системы (примеры в области БД и Веб)
- Различная степень *гранулярности* описания ресурса.

Основные свойства метаданных - 2

- Метаданные горизонтальной сферы *(«универсальные»)* / вертикальной сферы *(«специализированные»)*
- *Автономные* (отчужденные от описываемого ресурса)/*встроенные*
- *Независимые/зависимые* от контента описываемых ресурсов
- *Системные/пользовательские* метаданные
- *Структурированные/неструктурированные/слабоструктурированные* метаданные
- *Статические/динамические* (например, схема БД и каталог ЭБ)
- *Формализованные/неформализованные* метаданные
- *Явно/неявно представленные* (например, HTML-разметка / семантика ссылки в научной публикации)
- *Многоуровневость* метаданных: метаданные – это тоже данные, для них могут быть метаданные. Отсюда термины: *мета-метаданные, мета-мета-метаданные*…. (MOF, DC).

Функции метаданных - 1

- Функции метаданных зависят от конкретной сферы и условий их использования
- Далеко не исчерпывающий список функций:
  - ✓ Обеспечение интероперабельности и повторного использования ресурсов
  - ✓ Обеспечение интеграции данных из множества источников
  - ✓ Описание предметной области ИС: *концептуальная схема, онтология*
  - ✓ Описание баз данных и других репозиториев структурированных данных, поддержка механизмов управления их ресурсами: *схемы БД*
  - ✓ Описание других источников данных - контент ЭБ, открытые архивы, веб-сайты: *каталоги ЭБ и веб-сайтов, репозитории метаданных ОА*
  - ✓ Описание отдельных информационных объектов - таблиц БД, веб-страниц, информационных объектов в ЭБ: *описание таблицы в схеме БД, разметка веб-страницы, каталожная запись MARC или другие дескрипторы, поисковый образ документа в дескрипторной ИПС.*

Функции метаданных - 2

- ✓ Описание семантики источника информации, отдельного информационного объекта или его фрагмента: *рубрики рубрикаторов научной информации, набор значений элементов DC, семантическая (в частности, онтологическая) аннотация ресурса или его фрагмента, разметка средствами микроформатов, семантическая аннотация веб-страницы*

  При *онтологическом аннотировании данных* онтология = *метаданные,* при *онтологическом аннотировании метаданных* онтология = *метаметаданные*
- ✓ Описание представления данных на разных уровнях информационной архитектуры: *внешняя, концептуальная и внутренние схемы БД, разметка XML-страницы (иерархия элементов документа и XSL-спецификация)*
- ✓ Идентификация описываемых ресурсов: *первичный ключ таблицы БД, атрибут ID в DTD XML-документа, URL и URI, координаты точки в ГИС, DOI, ISBN, ISSN, штрих-код*
- ✓ Обеспечение функций управления данными БД и других источников информационных ресурсов
- ✓ Поддержка функций поиска информационных ресурсов.

Функции метаданных - 3

- ✓ Верификация данных на основе описаний структуры и ограничений целостности: *схема структурированных данных, DTD или XML-схема*
- ✓ Описание для пользователей свойств, назначения и других характеристик ресурсов (обычно на естественном языке)
- ✓ Описание ограничений доступа к информационным ресурсам
- ✓ Описание профилей пользователей: *полномочия, информационные потребности и пр.*
- ✓ Организация распространения информационных ресурсов: *на основе описаний ресурсов и информационных потребностей пользователей*
- ✓ Тематическая систематизация коллекций информационных ресурсов: *на основе рубрикаторов, тематических каталогов*
- ✓ Описание авторских прав на интеллектуальную собственность
- ✓ Использование для наукометрии в ЭБ: *семантика связей, рубрикаторы.*

Классификация метаданных

- Возможные классификации метаданных:
  - ✓ по их функциям
  - ✓ по уровням абстрактности
  - ✓ по их свойствам
  - ✓ по многим другим критериям
- Популярна агрегированная функциональная классификация:
  - ✓ *описательные:* контент ресурса, библиографические данные, аннотация, идентификаторы (URI, DOI, УДК…)
  - ✓ *структурные:* общая структура ресурса, ее компоненты (часть схемы базы данных)
  - ✓ *административные:* даты создания, обновления, владелец, полномочия пользователей…
- Имеются ее расширенные версии и модификации
- Оценка этих классификаций:
  - ✓ недостаточно строго определены
  - ✓ нет ясности в ее назначении
  - ✓ поэтому эти классификации в малой степени полезны.

Средства представления метаданных

- *Естественные языки* - наиболее содержательны, но не обеспечивают строгости, однозначности интерпретации, сложность компьютерной обработки: аннотации публикаций, сведения об авторах, об описываемых ресурсах, о содержании ресурсов
- *Искусственные языки* – большой пласт языков различного рода:
  - ✓ *описательные языки* с полным набором лингвистических элементов (алфавит, ситаксис, семантика): дескриптивное подмножество SQL, ODL, IDL CORBA, OWL, RDF, XML-Schema…
  - ✓ *языки разметки*: Tex, LaTex, SGML, HTML, XML, микроформаты…
  - ✓ *схемы метаданных* (наборы элементов метаданных): Dublin Core
  - ✓ *визуальные языки*: UML, ER-диаграммы, SADT (*Structured Analysis and Design Technique)*, семейство IDEF
- *Средства среды представления описываемых объектов :*
  - ✓ аудио, видео, специальные алфавиты…

Стандартизация метаданных

- *Стандартизация метаданных* – основа интероперабельности и повторного использования метаданных и описываемых ресурсов
- *Деятельность по стандартизации*: официальные органы, индустриальные компании и консорциумы, профессиональные сообщества
- Разработано большое число стандартов метаданных «*горизонтальной» и «вертикальной» сферы*
- Примеры стандартов *первой группы («горизонтальная» сфера)*:
  - ✓ дескриптивный подъязык языка SQL
  - ✓ язык описания объектов ODL консорциума ODMG
  - ✓ Open Information Model (OIM) консорциума Metadata Coalition
  - ✓ стандарты OMG: UML, CORBA IDL, MOF, Common Warehouse Model (CMW)
  - ✓ стандарты W3C: XML, XML Schema, RDF, RDFS, OWL, OWL2, WSDL
  - ✓ DCMI, NIST, ISO: Dublin Core (DC)
  - ✓ языки описания бизнес-процессов: BPEL, BPML
  - ✓ стандарты микроформатов
- Во второй группе значительное место принадлежит стандартам научных метаданных, созданным во многих областях исследований.

И снова об определении термина

- Лаконичное и качественное определение этого термина дать трудно
- Если стремиться к лаконичности, то учитывая разнообразие объектов, для которых создаются метаданные, возможно такое определение:

  Метаданные (метаинформация) объекта ИТ *= представленное с помощью какого-либо выразительного средства описание или фрагмент этого объекта, характеризующие его свойства.*
- Примеры:
  - ✓ *Метаданные - описания объектов:*
    схема базы данных, RDF-спецификация, набор значений элементов DC
  - ✓ *Метаданные - фрагменты объектов:*
    название статьи, фамилия автора, значение ключа строки таблицы БД, фрагмент мелодии («Угадай мелодию»), цитата из текста (при поиске содержащего ее текста), фотография фрагмента архитектурного сооружения (при поиске нужных сооружений), кадр из видео …
- Фрагменты объекта, используемые как его метаданные, выполняют функции идентификации этого объекта (например, *название статьи, значение первичного ключа*), характеристики его содержания (например, *аннотация*) и др.

Заключение

- Мир метаданных очень богат
- К сожалению, его богатство не осознано многими специалистами
- Перспективы: усиливается роль семантических метаданных
- Появление новых технологий и новых сфер применения по необходимости будет рождать новые системы метаданных
- Рассмотренные в докладе функции и свойства будут присущи и новым их видам.

Благодарю за внимание