Знания-Онтологии-Теории (ЗОНТ-13)

Онтология Электронных Документов Как Основа Для Извлечения Знаний О Предметной Области При Ее Моделировании

В.В. Ланин, Л.Н. Лядова

НИУ ВШЭ – Пермь, ул. Студенческая, д. 38, г. Пермь, 614070, Россия

lanin@perm.ru, llyadova@hse.ru

Аннотация. *Описан онтологический ресурс, представляющий различные аспекты жизненного цикла электронных документов в информационных системах. Предложенная модель документа позволяет унифицировать подходы к решению широкого спектра задач, возникающих в процессе обработки документов. Анализ показывает, что существующие на данный момент решения фокусируются чаще всего на решении какой-либо отдельной задачи, опираясь на соответствующее описание той или иной стороны документа. В предлагаемом подходе документ описывается в трех аспектах: формат (представление) документа, тип документа (выполняемые документом функции и состав реквизитов) и его структура. Работа состоит из двух частей, первая из которых посвящена обзору существующих онтологических ресурсов, так или иначе описывающих электронные документы, во второй части описываются компоненты предлагаемой онтологии электронных документов.*

Ключевые слова: электронный документ, формат документов, структура документа, онтология

1. Введение

При автоматической обработке электронных документов в информационных системах (ИС) возникает необходимость представления дополнительной информации о документе: его формате, виде и структуре, а также о представлении метаданных, описывающих дополнительные свойства документа [7]. В настоящее время данная информация рассредоточена (хранится как в самом документе, так в базах данных ИС, обрабатывающих документы) и специфична для каждой из задач, решаемых в течение жизненного цикла документа в ИС. Следствием этого является необходимость разработки единого механизма представления информации о документе. Решением может стать онтологический ресурс, описывающий различные аспекты электронного документа, рассматриваемые в течение всего его жизненного цикла. Этот ресурс может стать основой для решения широкого спектра задач, связанных с обработкой электронных документов в ИС.

2. Постановка задачи создания онтологического ресурса

Для решения задач обработки электронных документов, в частности, их поиска, анализа и классификации, каталогизации и эффективного хранения, генерации и поддержки их жизненного цикла необходимо иметь консолидированные знания об их структуре и содержании.

Необходима информация о следующих аспектах электронных документов:

- формат электронного документа;
- тип электронного документа;
- структура электронного документа.

При создании онтологического ресурса в него включаются понятия, относящиеся ко всем трём выделенным аспектам представления информации о документах. Каждый из них описывается онтологией. Понятия из различных аспектов должны быть связаны между собой. Таким образом, создаётся единая онтология электронных документов. Кроме того, ресурс должен поддерживать возможность расширения и уточнения для настройки на решение специфичных задач, возникающих при обработке документов в различных ИС в течение всего их жизненного цикла.

3. Обзор существующих решений для описания документов

3.1. Дублинское ядро метаданных

Дублинское ядро (Dublin core) [1] – набор элементов метаданных, предназначенный для описания контента документов различного типа (публикации, аудиозаписи, видеозаписи). Спецификация этого набора имеет статус официального международного стандарта (ISO: 15836-2003). Стандарт разделён на два уровня: *простой* (неквалифицированный, *simple*), состоящий из 15 элементов, и *компетентный* (квалифицированный, *qualified*), добавляющий к простому набору 3 дополнительных элемента и так называемые *тонкости* (или квалификаторы), которые уточняют семантику элементов. Особенностью Дублинского ядра является то, что каждый элемент опционален и может повторяться.

Дублинское ядро является мощным инструментом при описании ресурсов различного характера. Его неоспоримым преимуществом является распространённость и гибкость. Однако оно ориентированно на описание реквизитов документа, т.е. информации, напрямую не относящейся к контенту документа. Другие аспекты электронного документа в данном случае невозможно описать.

3.2. Онтологии форматов документов проекта «docOnto»

Разрабатываемые германской исследовательской группой KWARC (Knowledge Adaptation and Reasoning for Content) онтологии, в отличие от многих других проектов, ориентированы на разработку формального описания структуры (онтология документов в формате CNXML) и семантики документов (онтология документов в формате OMDoc). Участниками исследовательской группы разрабатываются также механизмы семантического индексирования документов и инструментальные средства обработки документов.

Онтология документов в формате CNXML (Connexions Markup Language, [4]) описывает такие понятия как параграф, раздел, ссылка и др. Онтология формализована на языке UML [5]. Представленная онтология довольно детально описывает структуру документа. К сожалению, активные работы по данному направлению приостановлены, последние изменения датируются 2007 г.

Второе направление в построении онтологий документов – описание семантики документов для узких предметных областей, где присутствуют хорошо формализованные документы, например, математические документы в формате OMDoc [10]. В онтологию включены математические понятия, логические связи между этапами доказательства теорем и многие другие понятия (см. рис. 1).

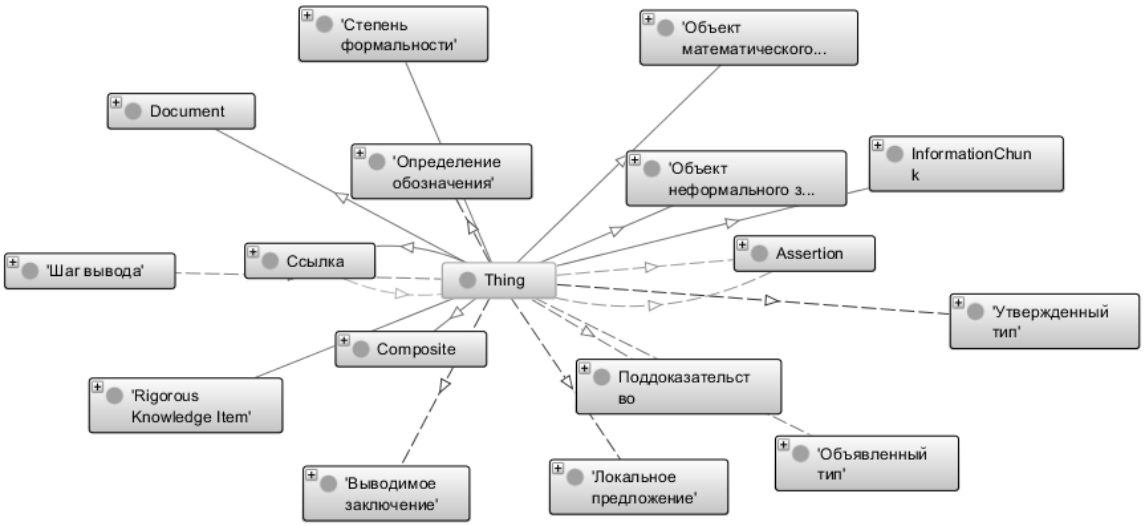

Рис.1. Фрагмент визуализации онтологии документа в формате OMDoc

3.3. Онтология документов проекта SHOE

Представленная ниже (рис. 2) онтология документов (Document Ontology) описывает большинство видов документов. Особое внимание уделяется печатным изданиям. Источниками разработки онтологии стали справочники Дублинского ядра [1] и классификатор документов PubMed.

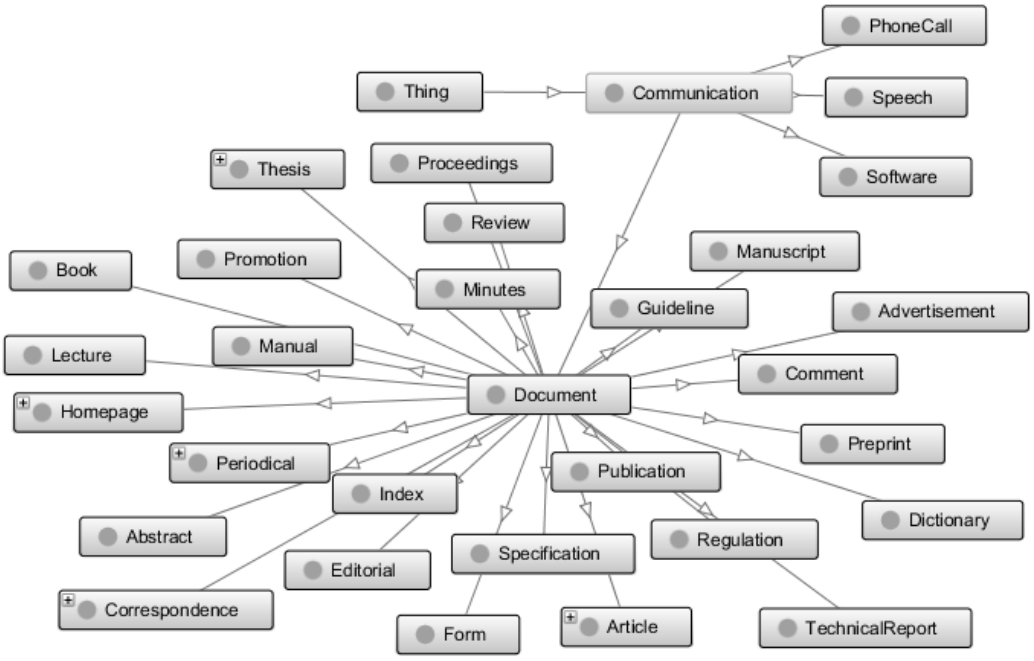

Рис. 2. Фрагмент визуализации онтологии документа проекта SHOE

3.4. Проект онтологии документов исследовательского центра Linked Data DERI

Онтология документов (см. рис. 3) разработана сотрудниками ирландского института DERI (Digital Enterprise Research Institute) и описана на языках RDFS и OWL-DL [9]. В онтологии представлены понятия, относящиеся к документации проектной деятельности. Разработчики

целенаправленно отказались от моделирования структуры и содержания документов в угоду гибкости и интероперабельности.

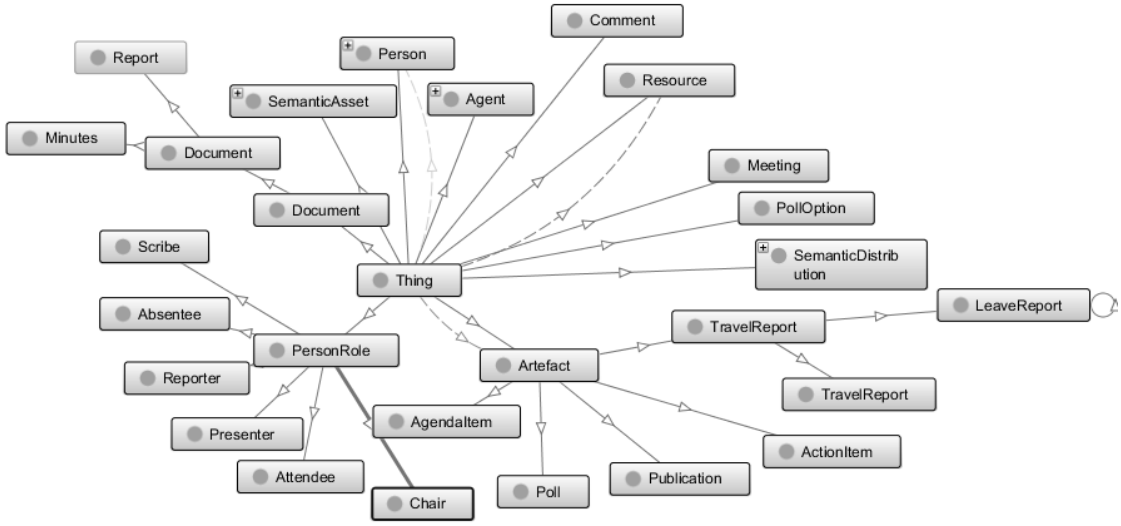

Рис. 3. Фрагмент онтологии документов DERI

3.5. Онтология документов проекта Muninn

Данная онтология (рис. 4) стала результатом проекта по обработке архивных документов первой мировой войны в рамках проекта Muninn WW1 [3]. Онтология описывает библиографические данные, происхождение и описание хранения цифровой копии. Большинство классов онтологии является дочерними классами FOAF. Такое решение принято для совместимости и, с другой стороны, делает возможным добавление специфики, связанной с обработкой документов, а именно поддержкой свойств для представления страниц документов, описания авторских прав и др. Одним из основных классов онтологии является *Document*, являющийся объединением классов *FOAF Document* и *Creative Commons Works*. Класс *Page* описывает страницы документа, цифровой образ страницы в свою очередь представлен классом *Image*.

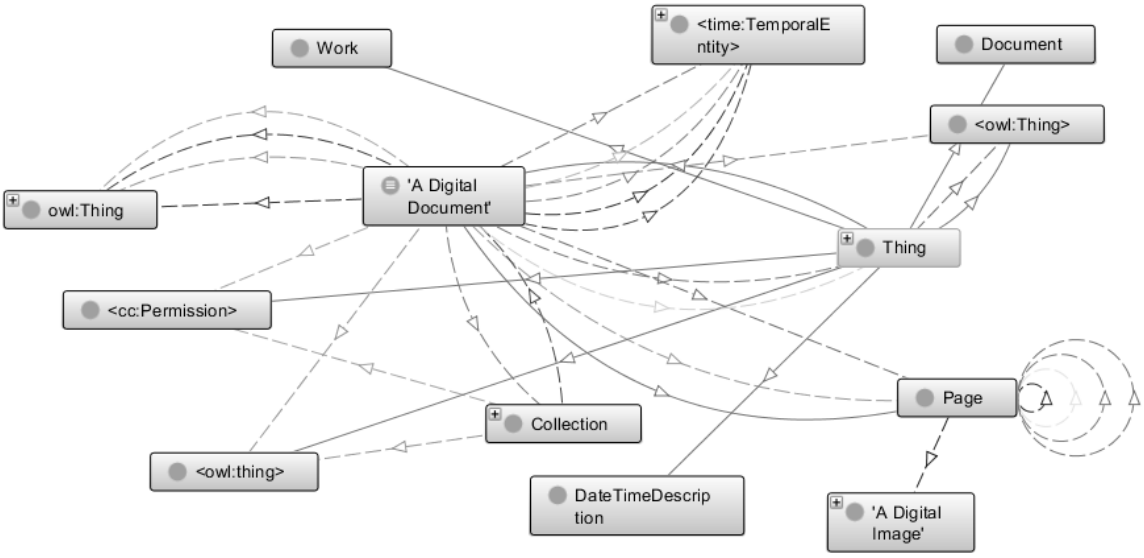

Рис. 4. Фрагмент визуализации онтологии документов проекта Muninn

Неоспоримым достоинством данной онтологии является описание различных аспектов документа, в частности структуры документа. Однако описание структуры ориентировано, прежде всего, на представление цифровых образов архивных документов.

4. Онтология электронных документов

Как отмечалось выше, для решения различных задач, связанных с обработкой документов, предлагается разработать единый онтологический ресурс. Основные компоненты этого ресурса описаны ниже.

4.1. Форматы электронных документов

В *онтологии форматов* можно выделить пять основных классов: «*Формат*», «*ПрограммныйПродукт*», «*Доступ*», «*НазначениеФормата*», «*ТипДанных*».

Класс «*ТипДанных*» имеет дочерние классы, соответствующие основным типам цифровых данных: библиотечные, звуковые, изобразительные, мультимедийные, текстовые, форматы архивов, форматы обмена и электронные таблицы и др. К классу «*Доступ*» относятся такие подклассы как «*открытый доступ*» и «*закрытый доступ*». Под *закрытым* форматом понимается то, что такие форматы являются собственностью конкретной организации. *Открытые* форматы – общедоступная спецификация хранения цифровых данных, свободная от лицензионных ограничений при использовании. В онтологию были добавлены также свойства «*hasFormat*», «*isFormatOf*», «*hasUseOf*» и «*isUsedIn*».

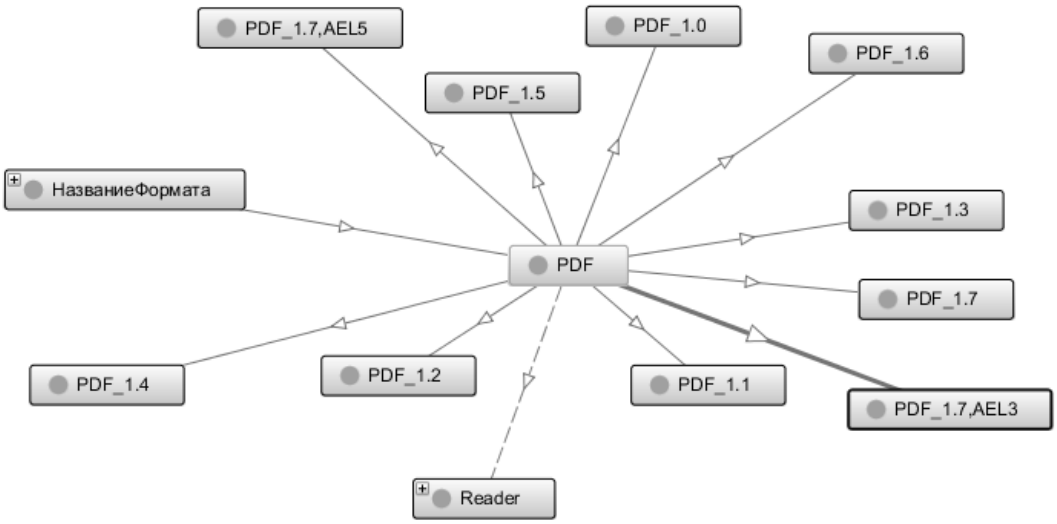

Рис. 5. Фрагмент визуализации онтологии документов: версии формата PDF

4.2. Типы документов

При проектировании *онтологии типов документов* за основу был взят ГОСТ Р 51141-98 [6]. В результате анализа были выделены следующие основные классы: реквизит; документ; содержание; принятие решения. Класс «*Реквизит*» содержит более 30 подклассов. Конкретные типы документов являются подклассами класса «*Документ*». Каждый вид документа обладает своим набором реквизитов и содержит определенную информацию. Класс «*Содержание*» включает в себя информацию о чертах и характеристиках, сгруппированных по схожести признаков, присущих каждому виду документа. В онтологию были включены свойства: «*имеет_реквизит*», два его подсвойства «*имеет_обязательный_реквизит*» и «*имеет_необязательный_реквизит*», а также обратное свойство «*является_реквизитом*». Кроме вышеперечисленных свойств в онтологию были добавлены следующие свойства: «*содержит*» для связи индивидов классов «*Документ*» и «*Содержание*», свойство «*издает*» для обозначения связи между классом «*Документ*» и «*Принятие_решений*».

Рис. 5. Фрагмент визуализации класса «*Докадная_записка*»

4.3. Структура документа

При моделировании структуры электронного документа (см рис. 6) предполагается, что документ представляет собой набор структурированных элементов, называемых фрагментами. Примерами фрагментов могут служить таблица, заголовок, реквизиты углового бланка и т.д. Фрагменты могут быть двух видов: элементарные фрагменты представляют простейшие неделимые элементы, такие как заголовок или дата составления документа, а составные фрагменты содержат в себе другие фрагменты. Фрагмент в свою очередь состоит из статистической и информационной частей. Статическая часть может быть представлена текстом, изображением, ссылкой, каким-либо специальным символом, кроме того, здесь может содержаться и информация для представления фрагмента. Информационная часть фрагмента либо указывает место для размещения элемента содержания, либо содержит множество фрагментов.

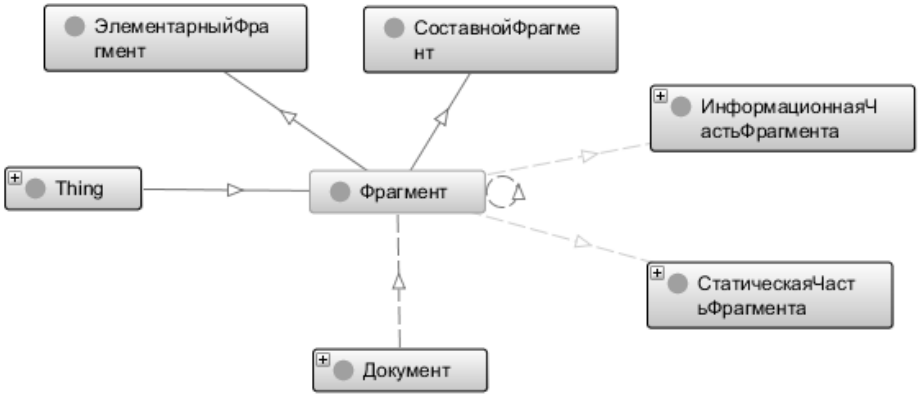

Рис. 6. Фрагмент визуализации онтологии структуры документа

Заключение

На данный момент в среде Protégé 4.3 реализованы прототипы компонентов онтологии. Следующий этап реализации ресурса – детализация онтологий и объединение разработанных компонентов в единый ресурс. Планируется также, что разработанная онтология будет использована в ряде проектов, связанных с разработкой предметно-ориентированных языков (Domain Specific Languages, DSL) для различных предметных областей на основе языкового инструментария MetaLanguage.

Благодарности

Работа выполнена при поддержке РФФИ (проект № 12-07-00763-а).

Литература

[1] Dublin Core Metadata Element Set, Version 1.1 [Электронный ресурс]. 2012. 14 июня. URL: http://dublincore.org/documents/dces/ (дата обращения: 18.06.3013).

[2] Document Ontology (draft) [Электронный ресурс]. URL: http://www.cs.umd.edu/projects/plus/SHOE/onts/docmnt1.0.html (дата обращения: 18.06.3013)

[3] Muninn Documents Ontology URL: http://rdf.muninn-project.org/ontologies/documents.html (дата обращения: 18.06.3013).

[4] XML Languages [Электронный ресурс]. URL: http://cnx.org/help/authoring/xml (дата обращения: 18.06.3013).

[5] CNXML/DocumentOntology [Электронный ресурс]. URL: http://mathweb.org/wiki/CNXML/DocumentOntology (дата обращения: 18.06.3013).

[6] ГОСТ Р 51141-98 Делопроизводство и архивное дело. Термины и определения. // Стандартинформ. М. 2006.

[7] *Ланин В*. Онтологии как основа функционирования систем обработки электронных документов // Материалы Всероссийской конференции с международным участием «Знания-Онтологии-Теории». Новосибирск, 2009, Т.2. С. 173-177.

[8] OMDoc/document ontology [Электронный ресурс]. URL: http://mathweb.org/wiki/OMDoc/document_ontology (дата обращения: 18.06.3013).

[9] *Varma P*. Project Documents Ontology [Электронный ресурс]. URL: http://vocab.deri.ie/pdo (дата обращения: 18.06.3013).

[10] OMDoc: Open Mathematical Documents [Электронный ресурс]. URL: https://trac.omdoc.org/OMDoc (дата обращения: 18.06.3013).