# Роли онтологий в электронной библиотеке КарНЦ РАН*

© В.А. Лебедев

Институт прикладных математических исследований КарНЦ РАН
Vl777@krc.karelia.ru

**Аннотация**

В статье рассмотрены проблемы построения онтологий научных дисциплин по описанию изученности природных объектов и систем, применение онтологий для систематизации и комплектации контента ЭБ КарНЦ РАН. Интеграция информационных ресурсов контента сопровождается их индексацией при помощи онтологий с целью последующего тематического поиска с использованием тех же онтологий для построения запросов.

## 1 Введение

Известный американский ученый Бернерс-Ли (Berners-Lee) в 2001 году предложил концепцию семантического Интернет (Semantic Web) как средство упорядочения контента сети и тем самым существенного уменьшения затрат ручного труда при поиске [12, 13]. Идея состоит в том, чтобы каждый информационный ресурс в сети сопровождала онтология, включаемая как значение специального атрибута в составе метаданных, например, в схеме метаданных Dublin Core это может быть атрибут Subject (тема) или Description (описание). В качестве информационного ресурса может выступать отдельный документ и/или коллекция документов. И каждый документ, включенный в коллекцию, и коллекция в целом сопровождаются онтологиями, причем каждая из них является фрагментом более общей онтологии предметной области, к которой относится публикуемый информационный ресурс. Имея открытую онтологию предметной области, возможно автоматически индексировать публикуемые ресурсы. Индекс при этом будет представлять фрагмент онтологии предметной области, т.е. будет онтологией документа. В этих условиях пользователь может применить для поиска программный агент, содержащий фрагмент онтологии предметной области и алгоритм для сравнения с ним онтологий ресурсов, найденных в сети. Тогда отклик на запрос практически не будет содержать информационного шума[1].

Очевидно, что для реализации Semantic Web в качестве первого шага требуется создание онтологий предметных областей, поэтому группа экспертов, работавшая по заданию Правительства РФ по определению перспективных направлений разработки информационно-коммуникационных технологий в России, определила как одно из приоритетных направление «Общедоступные методы и программные средства построения русскоязычных схем систематизации контента (программирование номенклатур, таксономий и онтологий предметных областей)» [10]. Под онтологией здесь понимается эксплицитная спецификация концептуализации предметной области, которая подразумевает использование некоторой математической модели и языка реализации этой спецификации [6, 14, 15].

Формально онтология определяется как **O = <X,R,F>**, где

- X — конечное множество понятий предметной области,
- R — конечное множество отношений между понятиями,
- F — конечное множество функций интерпретации [7].

Как видно, по форме – это определение графа с помеченными вершинами, где X – множество вершин, R – множество дуг, F- множество помет. Значение пометы интерпретируется некоторой функцией (функциями).

Будем понимать онтологию как иерархический граф связей терминов и названий, принятых в предметных областях, с их толкованиями, пометами и функциями интерпретации помет [11]. Основание: объекты, подлежащие описанию в ЭБ, обладают иерархической структурой, множество их свойств изучает комплекс научных дисциплин, иерархически соподчиненных, что и отображается в графе связей терминов. В составе ЭБ КарНЦ РАН [3-5] онтологии будут выполнять следующие роли:

- моделировать и представлять контент ЭБ,
- способствовать созданию контента, удовлетворяющего требованиям актуальности, достоверности и полноты,
- обеспечивать автоматическую индексацию электронных публикаций и
- построение точных запросов на поиск.

**Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Таким образом, необходимо разработать технологию построения и построить онтологии по биологии и науках о Земле, технологию формирования контента ЭБ с использованием онтологий, отработать технологию поиска релевантных документов при помощи онтологий.

Ряд функций указанных технологий нами был осуществлен ранее и опубликован в серии докладов на российских конференциях [1, 2, 8, 9]. В докладе излагаются решения и схемы технологий новых функций.

## 2 Построение онтологий

Для обеспечения достоверности, актуальности и полноты предметных онтологий необходимо разработать методологию их создания, соответствующую целям ЭБ КарНЦ РАН.

В качестве основы методологии принята следующая парадигма[2].

Природа Карелии состоит из объектов (предметов), подразделяемых на классы в соответствии с классификацией наук и научных дисциплин (например, по рубрикатору ГРНТИ). Каждый класс объектов характеризуется некоторым набором свойств (атрибутов), принимающих значения из соответствующих областей значений (доменов). Некоторые подмножества свойств объявляются признаками и используются непосредственно или их значения для различных классификаций объектов внутри класса. Множество свойств объектов разбивается на группы (темы), изучение которых является предметом соответствующей научной дисциплины.

Каждый объект вступает во взаимодействие с другими объектами, что является основой для выделения различных систем и подсистем. В системах объекты выполняют некоторые роли (функции), которые могут иметь различные оценочные названия (враги, союзники и т. п.), или выражаются соответствующими формулами.

Каждый объект любого класса обладает некоторым строением, то есть состоит из набора частей (тоже объектов), вступающих во взаимодействия и является системой (агрегатом).

Выделяют внешнее строение (морфологию) и внутреннее (анатомию).

Взаимодействие объектов в системах в некотором масштабе времени может быть неизменным (статика) или меняющимся (динамика). Подразделение взаимодействий объектов на классы и виды определяется в соответствующих научных дисциплинах. Статика определяет устойчивость, а динамика (процессы) – внешнее поведение (этологию), внутреннее функционирование (физиологию), происхождение, становление (генетика, генезис). Термины в скобках понимаются расширительно, в предметных областях конкретизируются и детализируются.

Методология построения предметных онтологий, основанная на данной парадигме, определяет структуру графа связей понятий (точнее, терминов и названий) предметной области.

Необходимо установить список классификаций и номенклатуры их классов. При необходимости зафиксировать соответствие (например, в виде табличной функции) значений признаков и классов. Установить номенклатуру (список) свойств (атрибутов) объектов класса, изучаемую данной научной дисциплиной. Затем определить их домены.

Некоторые термины являются многословными сочетаниями (например, сухие сосновые и смешанные леса, сырые засфагненные луга). Такого рода термины будем трактовать как конкатенацию названий классов различных независимых классификаций. В тех же примерах: леса, луга – типы растительности; сухие, сырые – классы по влажности; сосновые, засфагненные – классы по преобладающим видам растений и т.д.

Для обеспечения удобства поиска такого рода классификации в составе таксономии разносятся по уровням иерархии. В запросе они представляются в виде конъюнктивной (например, сырые $\wedge$ засфагненные $\wedge$ луга) или конъюнктивно-дизъюнктивной формы (например, сухие $\wedge$ (сосновые $\vee$ смешанные) $\wedge$ леса .

Далее следует устанавливать термины и названия, относящиеся к морфологии, анатомии, этологии и физиологии, то есть зафиксировать номенклатуры названий частей объектов и систем, их функции и оценки. При этом учитываются следующие типы отношений: классификации, агрегации, синонимии и полисемии. Технология, реализующая указанную методологию, состоит в следующем:

- Корневые понятия (термины) предметных областей принимаются по рубрикатору ГРНТИ.
- Начиная с корневых понятий, организуем поиск их значений (толкований) в Интернет или словарях.
- Используя найденное толкование, выделяем в нем термины более детальных понятий и ищем их толкования.
- Поступаем аналогично с терминами следующего уровня. И так до уровня значений свойств.
- В процессе поиска и нахождения терминов и их толкований фиксируем термин и URL статьи с наиболее полным толкованием его значения в связи с термином предыдущего уровня.


Рис. 1.

| Метка | Предок | Потомок |
|---|---|---|
| Биология А | Экология | Сообщества(экосистемы.биоценозы) |
| | | Связи |
| | | Виды |
| | | Популяции |
| | | Охрана окружающей среды |
| | Сообщества | Суша |
| | | Пресные воды |
| | | Моря |
| | | Атмосфера |
| Сообщества суши К | Типы | Биосфера |
| | | Зона |
| | | Подзона |
| | | Район |
| | | Ландшафт |
| | | Биогеоценоз |
| | | Местообитание(биотоп) |
| Сообщества суши К | Зона | Арктическая пустыня |
| | | Тундра |
| | | Лесотундра |
| | | Лес |
| Сообщества суши К | Подзона | Северная тайга |
| | | Средняя тайга |
| | | Южная тайга |
| | | Широколиственные леса |

Рис. 2.

- Таким образом, определяем как номенклатуру терминов и их связи, так и адреса (URL) толкований.
- После этого материалы передаются на экспертизу специалистам-предметникам, и по результатам экспертизы итерационно выполняется построение таксономий и механизма ссылок на толкования терминов.

В целом таксономия онтологии будет иметь структуру иерархического графа (древовидного или с полуциклами), фрагмент которого представлен на рис. 1.

Вершины графа – термины, дуги – отношения между ними (классификации и агрегации), отношения помечаются в узле разветвления. Отношения синонимии выделены в отдельную структуру (словарь). Полисемия (то есть наличие одинаковых по написанию терминов) разрешается в виду того, что такие термины могут находиться только в разных частях структуры.

Реализация таксономии представляется в виде таблицы (рис. 2), точнее, базы данных реляционной или объектной. Технология загрузки и редактирования таксономии и словаря синонимов отработана. Ведется разработка онтологий первой очереди по геологии, водным ресурсам, ботанике, зоологии, почвоведению, экологии, лесоведению и лесоводству. В настоящее время онтологии содержат около 2000 терминов, не считая видовых названий растений, минералов, химических соединений.

## 3 Технология формирования контента

На первых этапах создания ЭБ контент формировался только с привлечением специалистов КарНЦ РАН. С появлением онтологических моделей предметных областей возникает возможность расширения контента с использованием материалов, опубликованных в Интернете. Для осуществления этой возможности

разработана технология интеграции сторонних материалов в контент ЭБ.

Сущность и назначение интеграции сторонних материалов состоит в том, чтобы обеспечить доступ к ним посредством поисковых сервисов, имеющихся в ЭБ. Это сервисы поиска 1) по названиям коллекций и их документов и 2) с использованием онтологий для формирования тематических запросов.

Каждая коллекция (собственная или внешняя) должна пройти процесс импортирования, который включает формирование: записи в списке коллекций и списка документов коллекции.

Записи в списке коллекций содержат название коллекции и ее URL в виде гиперссылки. Аналогично записи в списке документов также содержат их названия и гиперссылку на текст документа. Для формирования этих списков имеется соответствующая технология [3, 4].

Отличие процесса регистрации интегрируемых сторонних коллекций заключается в том, что их документы могут быть представлены в различных форматах (HTML, PDF и др.) и могут не содержать списка документов в явном виде. Таким образом, необходимо будет разработать дополнительные технологические средства для формирования списков документов привлекаемых коллекций, аналогично тому, как формируются списки терминов онтологий и их толкований.

Для обеспечения тематического поиска документов в коллекциях производится их индексация с использованием соответствующей предметной онтологии. Структура индексного файла – это таблица, которая содержит имя документа, его URL и список встречающихся в его тексте терминов в порядке их иерархии и связей в онтологии.

Документы ЭБ по степени структуризации можно разделить на три категории: базы данных (таблицы), слабо структурированные (XML - документы) и неструктурированные (статьи в форматах PDF, HTML и т.п.). Таблицы и XML-документы структурно соответствуют структуре онтологии, поэтому процесс их индексации сравнительно прост. Документ прочитывается пословно. При этом рубрики документа соответствуют рубрикам онтологий, что обеспечивает сохранение порядка терминов в индексе, принятому в онтологии. Это важно для последующего поиска релевантных.

Неструктурированные документы могут содержать термины не в порядке их подчиненности в онтологии. Тогда, если не принять особых мер при их индексации, индекс документа будет содержать список терминов в порядке их нахождения в тексте, а не в порядке, принятом в онтологии, что впоследствии будет порождать информационный шум.

Чтобы избежать этого, в тексте документа в процессе его чтения сначала ищутся термины, близкие к корню онтологии. И если найден один, то дальше ищутся термины, подчиненные ему вплоть до листьевых терминов, и они помещаются в индекс. Далее ищется следующий термин корневых уровней и подчиненные ему и т. д. В результате список терминов в индексе будет иметь порядок, соответствующий онтологии.

После выполнения указанных операций сторонние коллекции считаются интегрированными в ЭБ и доступ к ним осуществляется при помощи сервисов нашей ЭБ.

Интеграция статей толкований терминов онтологий отличается тем, что списки статей, относящихся к данной предметной онтологии, включаются в соответствующий индексный файл. Причем индексация статьи может не производиться.

Реальные объекты, описываемые в документах коллекций, вступают между собой в различные отношения, которые указываются в виде их ролей в составе системы. Целесообразно использовать эту информацию для создания гиперссылок между документами. В результате получаем не просто наборы документов, а комплексы связанных документов, что полезно при изучении. Для решения этой задачи разработана соответствующая технология [2]. В итоге получим распределенную библиотеку, содержащую описание классов объектов Карелии и толкования терминов онтологии.

## 4 Поиск в ЭБ релевантных документов

Преимущество в использовании онтологий для формирования запросов на поиск заключается в том, что запрос в этом случае представляет собой фрагмент таксономии, в котором термины связаны в иерархию. Тем самым запрос уже не является простым списком терминов, а отражает их зависимость. При этом устраняется возможная полисемия терминов и тем самым отсекается значительная часть информационного шума в отклике на запрос.

Ранее нами была предложена обобщенная схема запросов [1], которая представляет собой редукцию предикатного выражения, а именно, нетерминалы в угловых скобках обозначают предикаты вида X=a, где X — слово в составе индекса, а — термин в запросе. С учетом объединения предметных онтологий на основе рубрикатора ГРНТИ схема запроса принимает следующий вид:

{ <Рубрика ГРНТИ><коллекция>}[<класс> ∧∨ …∧][<агрегат (тема)> ∧∨ … ∧]
[<характеристика> ∧][<список значений> ∧∨]… ∧∨
[<характеристика M > ∧][<список значений> ∧∨][<тема>∧∨ … ∧]
[<характеристика N > ∧][<список значений> ∧∨]… ∧∨

| Рубрика ГРНТИ | Коллекция | Тема | Характеристика 1 | Список значений 1 |
|---|---|---|---|---|
| Биология⋁ Ботаника | Сосудистые растения ⋀ | Экология растений ⋀ | Местообитания ⋀ | Сухие⋀Смешанные⋀ Леса⋁Поляны |

| Характеристика 2 | Список значений 2 |
|---|---|
| Хозяйственное значение ⋀ | Лекарственное ⋀ Противовоспалительное ⋀Пищевое⋀ (Ягоды ⋁Орехи) |

Рис. 3.

[<характеристика N + К > ⋀][<список значений ⋀ /⋁>]
[<класс> ⋀/⋁ … ⋀][<тема> ⋀/⋁ … ⋀][<характеристика> ⋀]
[<список значений> ⋀/⋁] … ,
где
<список значений> : = <значение> ⋀ /⋁<список значений>.

Наличие квадратных скобок указывает на возможность формирования и самых простых запросов из одного термина. При использовании дизъюнкций в составе фрагментов, заключенных в квадратные скобки необходимо правильно расставить круглые скобки, чтобы учитывать приоритеты логических операций.

Нетерминалы «Рубрика ГНТИ» и «коллекция», заключенные в фигурные скобки, определяют как раздел онтологии, так и коллекцию документов, в которой должен выполняться поиск. Очевидно, что поиск осуществляется в одной коллекции. При необходимости поиска в большем их числе запросы должны повторяться.

Нетерминалы «класс», «тема», «характеристика», «значение» отражают иерархическую структуру онтологии. При этом нетерминалы «класс» и «тема» подразумевают возможность иерархических классификаций.

Пример формирования запроса показан на рис. 3.

Очевидно, что построение запроса по указанной схеме непростая задача, поэтому предусмотрены средства оказания помощи пользователю в составлении запроса. Сначала он в процессе поиска в глубину по онтологии формирует список терминов для помещения в запрос, и только далее, обдумав свои потребности, переносит термины в запрос в порядке, определяемом иерархией и советами инструкций. При этом расстановка знаков конъюнкции и дизъюнкции и формирование оператора Select выполняется при помощи программы сервиса, которая контролирует допустимость конъюнктивных связей между терминами, как это показано ниже.

В нашем случае онтология представляет собой множество терминов предметной области, связанных между собой отношениями классификации, агрегации и синонимии.

Классификации разбивают некоторые исходные множества на группу непересекающихся подмножеств (классов) по определенным основаниям, в качестве которых могут использоваться наличие или отсутствие у объекта определенных атрибутов (признаков) и/или определенных значений атрибутов. Классификации могут быть одноуровневыми или многоуровневыми (иерархическими), многоуровневость производит последовательная классификация сначала исходного множества, а затем его подмножеств, подмножеств этих подмножеств и т.д. Одно и то же множество может быть классифицировано несколько раз с использованием различных оснований. По определению, в классификациях допускаются конъюнкции между терминами, лежащими на одном пути в графе онтологии. Все остальные конъюнкции являются пустыми, т. к. связывают непересекающиеся подмножества. В этих условиях, чтобы проверить допустимость любой конъюнкции, заданной в запросе, достаточно проверить, лежат ли входящие в нее термины на одном пути в онтологии и, если не лежат, сообщить об этом пользователю, чтобы он исключил эти конъюнкции из запроса.

Агрегации в отличие от классификаций позволяют представить класс объектов в виде совокупности частей или свойств. Отдельные объекты класса описываются указанием значений свойств или порядковых номеров частей (присваиваемых в процессе производства). При поиске объектов класса в коллекциях в этих случаях допускаются конъюнкции между названиями свойств или частей. При поиске конкретных объектов нужно указывать также значения свойств или номера частей. Запись конъюнкции должна быть аналогичной предыдущей схеме.

В данном случае допустимые конструкции в запросах для классификаций и агрегаций вступают в противоречие. Для его разрешения достаточно пометить в онтологии классификации и агрегации различными знаками (например, классификации метятся буквой К, а агрегации – буквой А) (см. рис. 2). Тогда упомянутый выше контроль допустимости конъюнкций достаточно дополнить анализатором этих пометок.

Синонимические гнезда терминов в онтологиях представляются отдельными словарями синонимов.

Когда пользователь помещает в запрос очередной термин, выполняется поиск в словаре синонимов, и если они есть, то автоматически в запрос помещается дизъюнкция всего синонимического гнезда. Тем самым предотвращается возможность включения в запрос пустых конъюнкций. Доработка и реализация алгоритмов индексирования документов выполнена Старковой В.Г..

## Литература


[1] Вдовицын В. Т., Лебедев В. А. Онтологии для тематического поиска данных в коллекциях электронной библиотеки. // Труды десятой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”. Дубна. 2008. С. 63-69.

[2] Вдовицын В. Т., Лебедев В. А., Брагин С. В., Старкова В. Г., Луговая Н. Б. Развитие сервисов электронной библиотеки научных информационных ресурсов //Труды Всероссийской научной конференции Научный сервис в сети Интернет: технологии параллельного программирования, г. Новороссийск, 24 – 29 сентября 2007 г. Издательство Московского университета. 2007. С. 305-310.

[3] Вдовицын В. Т., Лебедев В. А., Луговая Н. Б., Сорокин А. Д., Старкова В. Г.. Развитие и разработка технологии публикации и поиска документов в электронных коллекциях // Труды Восьмой Всероссийской научной конференции по электронным библиотекам, Суздаль, 2006. С. 162-167.

[4] Вдовицын В. Т., Сорокин А. Д., Луговая Н. Б.. Развитие программных сервисов и контента ЭБ КарНЦ РАН. // Труды Седьмой Всероссийской научной конференции по электронным библиотекам, Ярославль, 2005. С. 92-97.

[5] Вдовицын В. Т., Сорокин А. Д., Луговая Н. Б.. Электронная библиотека научных информационных ресурсов КарНЦ РАН. // Труды Шестой Всероссийской научной конференции по электронным библиотекам, Пущино, 2004. С. 41-46.

[6] Добров Б. В., Лукашевич Н. В. и др. Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Труды Седьмой Всероссийской научной конференции по электронным библиотекам, Ярославль, 2005. С. 70-76.

[7] Загорулько Ю. А. Методы и метологии разработки, сопровождения и реинжиниринга онтологий. Онтологическое моделирование. Труды Симпозиума. Звенигород, май 2008. С. 167-200.

[8] Лебедев В. А., Старкова В. Г., Брагин С. В. Представление онтологии научной коллекции «Водные ресурсы региона» // Труды шестой Всероссийской конференции по электронным библиотекам. Пущино, 2004. С. 86-92.

[9] Лебедев В. А., Старкова В. Г., Брагин С. В. Применение онтологии для ведения и доступа к данным коллекции «Природные ресурсы региона». // Труды седьмой Всероссийской конференции по электронным библиотекам». Ярославль, 2005. С. 87-91.

[10] Перспективные направления развития российской отрасли информационно-телекоммуникационных технологий (Долгосрочный технологический прогноз Российской IT — Foresight) М. , 2007. 223 с.

[11] Фазлиев А. З. Рассуждения о понятии “онтология”. Онтологическое моделирование. Труды Симпозиума. Звенигород, май 2008. С.278-296.

[12] Хорошевский В. Ф. Онтологические модели и Semantic Web: откуда и куда мы идем? Онтологическое моделирование. Труды Симпозиума. Звенигород, май 2008. С. 13-45.

[13] Berners-Lee T., Hendler J., Lassila O. The Semantic Web. Scienific American. 2001.

[14] Gruber T. R. A Translation Approach to Portable Ontology specification // Knowledge Acquition, N 5, 1993.

[15] Uschold M., Gruninger M. Ontologies: Principles, Methods and Applications. // Knowledge Engineering Review, N 11, 1996.


## Roles of ontologies in Karelian Research Centre's digital library


Lebedev V. A.
Institute of Applied Mathematical Research,
Russian Academy of Sciences



The paper considers problems of building ontologies of scientific disciplines by descriptions of the degree of coverage of natural objects and systems by studies, application of the ontologies to systematization and compilation of the contents of the Karelian Research Centre's digital library. Integration of the information resources in the contents is accompanied by their indexing through ontologies to enable further thematic search using the same ontologies to build queries.



[*] Работа поддержана грантом РФФИ № 08-07-00085а.


[1] Информационный шум – документы, ошибочно включенные в состав отклика на запрос из-за содержания в них ряда терминов, обладающих полисемией в различных предметных областях.

[2] Парадигма – это наиболее общая картина устройства природы, в данном случае – описания изученности природных объектов и систем.