УДК 004.9:81’374.82

Е.В.ПОТАПОВА

Таврический национальный университет им.В.И Вернадского, Симферополь

ИНСТРУМЕНТАЛЬНАЯ СИСТЕМА ДЛЯ СОЗДАНИЯ МНОГОЯЗЫЧНОЙ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

Анотація. У статті подано концептуальну модель багатомовної онтології з Фізики магнітних явищ як лексикографічної системи особливого типу. На її підставі розроблено інструментальну система для створення і управління багатомовною онтологією на основі засобів і методів СУБД. Програмний інтерфейс і алгоритм опису термінів не залежать від предметної галузі й можуть бути використані для інших предметних галузей.

Ключові слова: програмний інтерфейс, багатомовна онтологія, лексикографічна система, електронний словник.

Аннотация. В статье представлена концептуальная модель многоязычной онтологии по Физике магнитных явлений как лексикографической системы особого типа. На основе данной модели разработана инструментальная система для создания и управления многоязычной онтологией на основе средств и методов СУБД.

Ключевые слова: программный интерфейс, многоязычная онтология, лексикографическая система, электронный словарь.

Abstract. The article presents a conceptual model for multilingual domain ontology (Physics of the magnetic phenomena) as lexicographic system of special type. On the base of this model the instrumental system for multilingual ontology creating and managing was designed. The instrumental system based on the tools and methods of the data base management systems. The program interface and the algorithm of the terms description does not depend on domain and they can be used to create the multilingual ontology for other domains.

Keywords: programming interface, multilingual ontology lexicographical system, electronic dictionary.

Введение

Создание многоязычных онтологий имеет практическое значение в области согласования терминологии в различны предметных областях (ПрО), развития технологий информационного поиска (организации многоязычного поиска), машинного перевода в рамках заданной предметной области и т.д. Одной из важных проблем в данном направлении является моделирование взаимодействия терминологических систем для нескольких (более двух) языков. Применение теории лексикографических систем [1, 2] к концептуальному моделированию онтологии предметной области «Физика магнитных явлений» как лексикографической системы особого типа позволило создать концептуальную модель многоязычной онтологии ПрО [3]. Предложенная модель реализована в виде лингвистической базы знаний по Физике магнитных явлений с представлением информации на русском, украинском и английском языках. В статье представлен опыт разработки инструментальной системы для создания и управления трехъязычной (русский-украинский-английский) лингвистической онтологией «Физика магнитных явлений» (ФМЯ). Концептуальная модель многоязычной онтологии, разработанная ранее [3, 4], здесь приводится кратко, для связности изложения.

Постановка цели и задачи научного исследования

Цель исследования: разработка методов описания предметных и лингвистических знаний на примере предметной области «Физика магнитных явлений» с помощью трехъязычной лингвистической онтологии, а также методов и инструментальных средств управления лингвистической онтологией.

Задачи исследования:

1) Создание инструментальной среды, реализующей предложенную концептуальную модель.
2) Разработка интерфейса настройки на определенную предметную область.
3) Обеспечение визуализации онтологии в виде комплекса словарных статей и когнитивных карт терминов в электронном виде.

Концептуальная модель

Формальная модель онтологии — это упорядоченная тройка конечных множеств [5]:

$$O = \langle T, R, F \rangle, \qquad (1)$$

где $\boldsymbol{T}$ — термины ПрО, $\boldsymbol{R}$ — отношения между терминами, $\boldsymbol{F}$ — функции интерпретации, заданные на терминах и/или отношениях онтологии ПрО.

В результате концептуального моделирования предметной области ФМЯ посредством лингвистической онтологии предложена оригинальная модель представления знаний ПрО с расширенным описанием ассоциативных – проблемно-специфических связей (подробнее в [6]). Для определения типов проблемно-специфических связей было введено понятие «лексико-онтологического класса» (ЛОК) представляющего собой некоторое нечеткое множество, исследованы особенности перехода терминов онтологии между классами и существование нечетких семантических состояний [4]. Для их описания в концептуальной онтологической модели ПрО ФМЯ использованы нечеткие множества Л. Заде и теория семантических состояний Колмогорова-Широкова [3]. Определена функция, описывающая нечеткое семантическое

состояние понятия онтологии. В формальную модель онтологии были введены параметры, отражающие данную ситуацию.

Таким образом, лексикографическая параметризация для любого термина $t \in \langle T \rangle$ из формулы (1) включает в себя следующие параметры:

$$t(t^{L1}, t^{L2}, t^{L3}, t^{OC} \langle t^G \rangle, t^D, \langle t^C \rangle, \langle t^I \rangle, \langle t^R \rangle), \quad (2)$$

где t^{L1} – орфографический стандарт (русский язык - $L1$), t^{L2} – переводной эквивалент на язык $L2$ (украинский язык), t^{L3} – переводной эквивалент на язык $L3$ (английский язык). t^{OC} - ЛОК термина, значение которого является элементом $\langle OntC \rangle$ - множества ЛОК установленных для данной ПрО. $\langle t^G \rangle$ - множество грамматических параметров, t^D – дефиниция текстовая, $\langle t^C \rangle$ – множество контекстных примеров, $\langle t^I \rangle$ - множество вспомогательной информации. $\langle t^R \rangle$ - множество связей с другими терминами - онтологическая окрестность термина.

Однако данная модель не разрешает проблему установления адекватного соответствия между переводными эквивалентами трех языков, так как в лексикографической параметризации термина онтологии параметры t^{L2} и t^{L3} могут содержать несколько значений. Для этого концептуальная модель онтологии была переработана в рамках теории лексикографических систем (Л-систем) [3].

В типологическом смысле Л-система – это информационный объект, сочетающий в себе черты модели данных, модели знаний и логико-лингвистического исчисления [1,2]. Основными системообразующими отношениями Л-системы являются: «субъект-объект» и «форма-содержание». Основным системообразующим инвариантом Л-системы является лексикографический эффект в информационных системах. Лингвистическая онтология предметной области отличается от других Л-систем тем, что помимо языкового уровня развития лексикографических эффектов, имеется надъязыковой уровень – уровень метаописаний понятий и их отношений. Такое разделение системы на два уровня позволяет создать модель для лингвистической онтологии предметной области на двух и более языках с общим ядром понятий и отношений между ними.

Преобразование концептуальной модели 3-х язычной онтологии ПрО в лексикографическую систему (LS) позволило выделить в ней два уровня: $LS \rightarrow \left(LS^{L1}, LS^{L2}, LS^{L3}\right)$ и определить взаимодействие уровней и компонентов полученной системы [3].

Уровень 1 в целом соответствует формальной метаонтологии понятий ПрО и содержит информацию о системе, не связанную с конкретным языком. Множество объектов *Ob(LS)*= $\{T, OntC, Pers, Br, R\}$, где *T*-множество терминологических понятий (метаописаний - индексов), *R* – множество типов связей между понятиями, *OntC* – множество ЛОК. Множество персоналий - *Pers* и множество разделов Физики магнитных явлений - *Br* соответствуют $\langle t^I \rangle$ - множеству дополнительной информации.

Множество отношений *RelOb(LS)*= $\{TT, TPers, TBr)$, где *TT* – множество связей между понятиями, что в онтологической модели соответствует $\langle t^R \rangle$. $TPers$ - множество связей между множеством понятий и множеством персоналий, TBr - множество связей между множеством понятий и множеством разделов ПрО.

Уровень 2 – отображение формальной онтологии понятий в языковые системы. Множество объектов $Ob(LS^{Li}) = \{T^{Li}, TCont^{Li}, TDef^{Li}\}$, где T^{Li} -множество терминов языка Li (в онтологической модели это множества t^{L1}, t^{L2}, t^{L3}); для $\forall t \in \langle T^{Li} \rangle$ $t\ (t^{Li}, \langle T^G \rangle, tDef^{Li})$, где t^{Li} - орфографический стандарт, T^G - множество грамматических характеристик термина ($\langle t^G \rangle$ в онт.м.). Текстовая дефиниция термина - $tDef^{Li}$ (t^D в онт.м.) и $\langle TCont^{Li} \rangle$ -множество контекстов термина на языке *Li* ($\langle t^C \rangle$ в онт.м.).

Множество отношений $RelOb\,(LS^{Li}) = \{TTCont^{Li}\}$ - это множество связей между множеством терминов и множеством контекстных примеров (коллекцией текстов).

Ограничение целостности заключается в том, что :

1) никакой элемент из T^{Li} уровня 2 не может существовать без связи с соответствующим элементом *T* уровня 1;

2) одному элементу *T* уровня 1 может соответствовать несколько элементов из T^{Li} уровня 2 в каждом из языков;

3) связь между объектами (терминами) различных языков (т.е. перевод) осуществляется только через элемент *T* уровня 1.

Таким образом, уровень 1 концептуальной модели содержит формальные данные онтологии, в то время как уровень 2 содержит лингвистическую часть информации о предметной области. Такое разделение позволяет проектировать многоязычную систему, для 2-х и более языков. Взаимодействие между комплексами лингвистических данных различных языков осуществляется только через формальную онтологию уровня 1.

Описанная концептуальная модель многоязычной лингвистической онтологии справедлива только для терминологических систем предметных областей и не используется для моделирования общеязыковой лингвистической онтологии для двух и более языков.

Формула словарной статьи

Процесс абстрагирования словарной (лексикографической) структуры представляет собой своеобразную расшифровку, реконструкцию того лексикографического эффекта (ЛЭ), который привел к образованию этой структуры. Вследствие того, что в лингвистической онтологии наблюдается два уровня ЛЭ, структура словарной (онтологической) статьи для некоторого понятия *T* выраженного термином T^{L1} в языке *L1* с грамматическими параметрами G^{L1} имеет вид :

$$T^{L1}(T, G^{L1}) = C^O + C^L;$$

$$C^O = OntC + TT + TPers + TBr;$$

$$C^L = S^{L1} + T^{L2} + T^{L3} + TDef^{L1} + TCont^{L1}, \text{ где } T^{L2} \supseteq S^{L2} \text{ и } T^{L3} \supseteq S^{L3},$$

где C^O - комплекс онтологических параметров, C^L - комплекс лингвистических параметров.

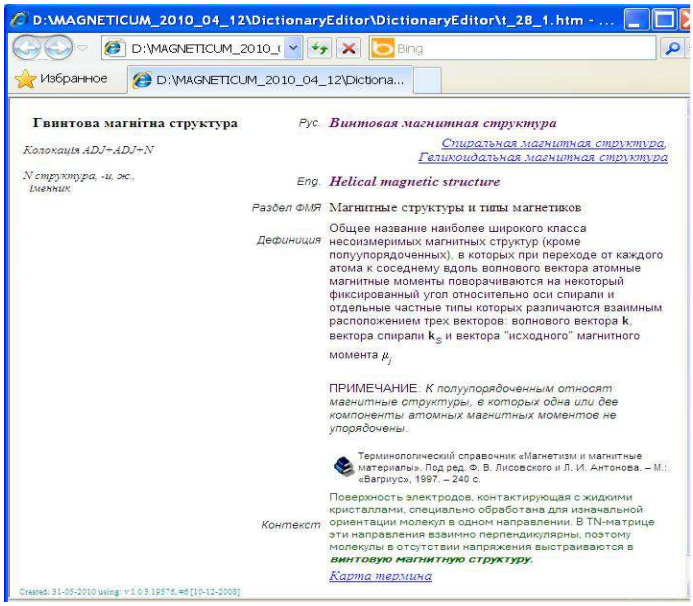

Рисунок 1 − Интерфейс пользователя. Словарная статья термина

Разделы переводных эквивалентов T^{L2} и T^{L3} на языки *L2* и *L3* соответственно – включают в себя весь синонимический ряд (классы условных эквивалентов S^{L1}, S^{L2}, S^{L3} соответственно). Остальные обозначения соответствуют вышеописанной концептуальной модели.

При отображении словарной статьи термина для пользователя в явном виде присутствуют лингвистические параметры. Из онтологических параметров отображаются только *TPers* и *TBr* – персоналии и раздел ФМЯ связанные с понятием, выраженным данным термином. Информация об онтологических связях термина отображается в виде когнитивной карты термина – фрагмента онтологической схемы (подробнее в [7]). На рис.1- 2 - пример словарной статьи и карта термина.

Рисунок 2 − Интерфейс пользователя. Пример когнитивной карты термина

Инструментальная система

Внутренняя модель. Исходя из двухуровневой концептуальной модели изложенной выше, в структуре онтологической лексикографической базы данных (ЛБД) данные, не соотносимые с конкретным языком (связи терминологических понятий, набор лексико-онтологических классов, персоналии, раздел ФМЯ) образуют отдельную группу таблиц. С другой стороны выделены три языковые группы таблиц: английская, русская и украинская (рис.3).

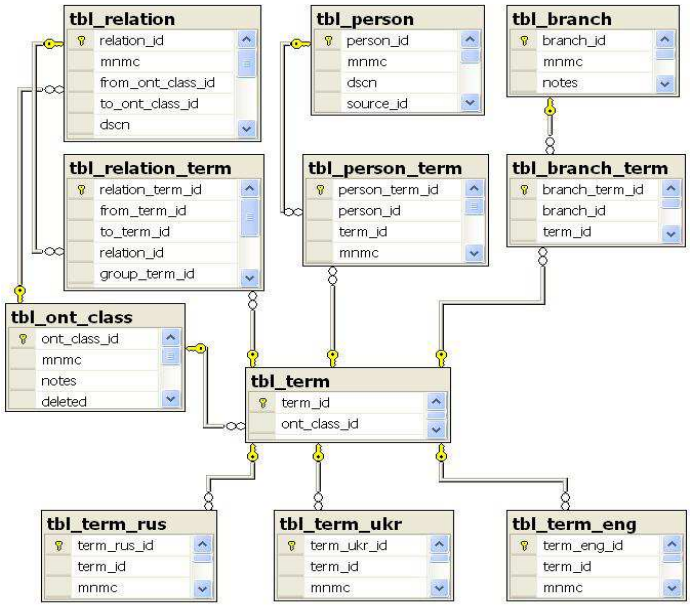

Рисунок 3 − Общая схема лексикографической базы данных

Главным реестром, связывающим все группы таблиц, является таблица tbl_term, которая содержит индексы терминологических понятий. Таблицы tbl_term, (tbl_relation_term, tbl_relation, tbl_ont_class – связи понятий и онтологические классы), (tbl_person_term, tbl_person – информация о персоналиях связанная с открытием или исследованием тех или иных физических явлений), (tbl_branch_term, tbl_ branch – перечень разделов физики магнитных явлений) соответствуют 1 уровню концептуальной модели – формальной метаонтологии. Все множества объектов хранятся в отдельных таблицах. Связь между множествами объектов (связь типа «многие-ко-многим») обеспечивается через таблицы составных (сложных) ключей tbl_relation_term, tbl_person_term, tbl_branch_term.

Второй уровень образуют языковые группы таблиц, связанные друг с другом через таблицу tbl_term. Рассмотрим на примере группы таблиц русского языка: tbl_term_rus – главный реестр терминов, каждый термин может иметь одно текстовое определение (tbl_term_rus_dscn) и несколько входов в коллекцию научных текстов (tbl_collection_rus). В то же время каждый текст из tbl_collection_rus может быть связан с несколькими терминами, что обеспечено сложным ключом tbl_term_collection_rus. Список библиографических источников содержится в таблице tbl_source.

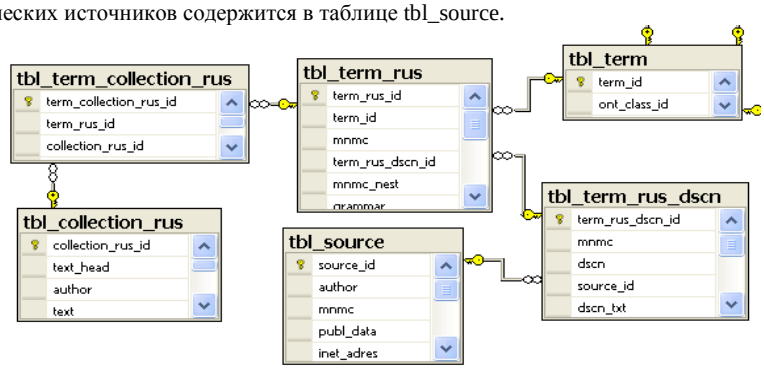

Рисунок 4 − Группа таблиц русского языка

Ввиду наличия формул и рисунков в текстовых описаниях терминов, в коллекции текстов и текстах представляющих персоналии, для хранения такой смешанной информации использован формат *rtf*.

Внешнее представление и инструментальная Л-система. Редактор онтологии реализует концептуальную модель и предоставляет возможность управления данными ЛБД представленными на Рис.3-4. Редактор реализован как локальное Windows-приложение (C#) с доступом к источнику данных (SQL Server).

Графический интерфейс редактора лексикографической базы разработан в соответствии с концептуальной моделью и структурой словарной статьи. Главное окно предоставляет пользователю выбор режима работы: 1)Онторедактор, 2)Редактор языковых реестров.

Функции **Онторедактора :**

1) ввод нового понятия (должен быть введен хотя бы один языковой эквивалент);
2) установка переводных соответствий (связь между языковыми реестрами);
3) установка связей понятия с другими понятиями (рис.5.);
4) установка связей с персоналиями, с разделом ФМЯ;
5) демонстрация онтологической схемы связей понятия (карты термина).

Функции **Редактора языковых реестров**:

- указание грамматических характеристик термина; модель коллокации; ввод условных языковых эквивалентов (синонимов);
- ввод текстовой дефиниции;
- установка связей с коллекцией текстов.

Кроме того здесь тоже есть вкладка «Связи» для поиска и установки связей термина в результате анализа текстовой дефиниции или текста из коллекции текстов. Поиск связи автоматизирован, тип связи определяет эксперт. Введенные данные отображаются в виде словарной статьи термина и его когнитивной карты (рис.1,2).

Для всех множеств объектов уровня 1 и уровня 2 концептуальной модели существуют отдельные окна для редактирования: «Редактор связей», «Онтологические классы», «Персоналии», «Разделы ФМЯ», «Коллекция текстов», «Библиографические источники».

Перечислим технические и процедурные особенности разработанной инструментальной системы и онтологической лексикографической базы данных:

1. Модульная структура: позволяет использовать отдельные библиотеки объектов и фрагменты (группы таблиц БД) как отдельные приложения или для модернизации существующих информационных систем. Т.е. разработанная инструментальная система позволяет разрабатывать информационные ресурсы следующих типов:
   - 3-х язычная онтология ПрО с лингвистическим компонентом описания информационных единиц и когнитивными картами терминов, совокупность, которых представляет собой графическое отображение онтологии ПрО;
   - Формальная онтология понятий (реестр понятий должен быть выражен на одном из языков) + карты терминов;
   - 3-х язычный терминологический словарь;
   - Словарь терминов + коллекция текстов, размеченная этими терминами (реализовано для русского языка).
2. Адаптация многоязычной лингвистической онтологии ПрО под определенную предметную область включает в себя, помимо создания словаря ПрО, создание классификации понятий и определение типов проблемно-специфических связей между понятиями на основе заданной классификации. Кроме того, инструментальная система позволяет настраивать таксономические связи.
3. Реализована функция определения онтологического окружения термина на основе теории нечетких семантических состояний и системы лексико-онтологических классов (подробнее в [4]). Результат может быть использован для уточнения поисковых запросов.
4. Доказательство и интроспекция БЗ обеспечивается структурой SQL запросов и другими возможностями СУБД.

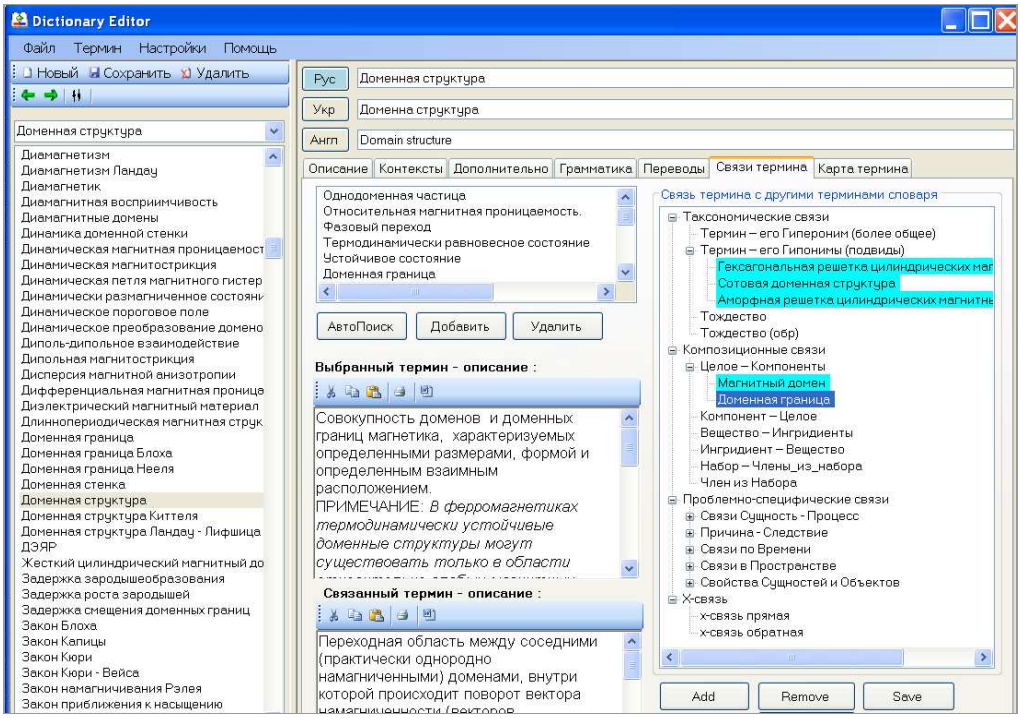

Рисунок 5 - Инструментальная система. Онторедактор. Вкладка «Связи термина»

Посредством разработанной инструментальной системы была проведена работа по установлению связей между понятиями в ряде разделов Физики магнитных явлений и получены соответствующие схемы связей – карты терминов, которые являются взаимосвязанными фрагментами онтологической схемы предметной области. На данном этапе существует 276 карт терминов. Карты терминов генерируются инструментальной системой в автоматическом режиме в формате HTML.

Пользовательская версия. На любом этапе редактирования ЛБД содержимое онтологической базы знаний может быть опубликовано в формате html для конечного пользователя в виде отдельного

электронного продукта «Словарь терминов по ФМЯ» на трех языках (русский-украинский-английский) с когнитивными картами терминов.

Заключение

Расширение онтологической модели ПрО «Физика магнитных явлений» средствами теории лексикографических систем позволило создать модель многоязычной лингвистической онтологии ПрО как лексикографической системы особого типа.

Реализация предложенной концептуальной модели многоязычной лингвистической онтологии ПрО в виде инструментальной среды интегрированной с СУБД позволяет успешно решать ряд задач управления онтологией (логический вывод, нахождение противоречий) и пополнения онтологии как вручную, так и путем экстракции терминов и их связей из текстов (реализовано для русского языка).

Список литературы

1. Широков В.А. Елементи лексикографії.– К.: Довіра, 2005. – 304с.
2. Широков В.А. Інформаційна теорія лексикографічних систем. – К.:Довіра, 1998. – 331с.
3. Широков В.А., Потапова Е.В. Онтология предметной области как лексикографическая система особого типа. //«Казанская наука», Казань. №12 2012, С.209-213.
4. Потапова Е.В. Модель лингвистической онтологии предметной области с нечеткими семантическими состояниями терминов. // Научно-технический журнал «Бионика интеллекта», Харьков, ХНУРЭ, № 2(79), 2012, с.95-102.
5. Добров Б.В., Соловьев В.Д., Лукашевич Н.В., Иванов В.В. Онтологии и тезаурусы. Модели, инструменты, приложения. Бином, 2009. - 173 с.
6. Бержанский В.Н., Потапова Е.В. Классификация связей между понятиями в онтологии по физике магнитных явлений. / MegaLing 2010: сб.научн. трудов – К.: Довіра, 2010, С.12-21.
7. Бержанский В.Н., Дикарева С.С., Потапова Е.В. Принципы конструирования когнитивных карт терминов в многоязычной лексикографической системе по физике магнитных явлений./ Актуальні проблеми прикладної лінгвістики. Матеріали міжнародної Інтернет-конференції (23-24 червня 2011). – Умань, 2011, С. 23-26.

Сведения об авторе

Потапова Елена Владимировна - м.н.с. Межведомственного центра когнитивной и прикладной лингвистики Научно-исследовательской части Таврического национального университета им.В.И Вернадского, e-mail: helen1pota@rambler.ru , Украина, АР Крым, 95050 Симферополь, ул.Киевская, д.147, кв. 51.