Развитие лексической онтологии по аналитической химии: разработка тезауруса в виде реляционной базы данных, миграция данных в MS WSS 3.0 и публикация в Интернете♣

© В.И Широкова, В.П. Колотов, М.В. Аленина

Институт геохимии и аналитической химии им. В.И.Вернадского РАН
shirokova@geokhi.ru

Аннотация

Разработана лексическая онтология в виде многоязычного тезауруса терминов по общим вопросам аналитической химии. Особенностью этого тезауруса является то, что он построен на реляционной модели взаимосвязи сущностей, терминология из разных официальных источников впервые интегрирована на семантической основе. Тезаурус включает наборы синонимов, родовидовые отношения терминов, симметричную ассоциацию понятий.

Для публикации тезауруса в Интернете в системе MS SharePoint Services (WSS 3.0) разработана система запросов в реляционной базе данных для компиляции сводной таблицы, способ отображения сводной таблицы в MS WSS 3.0 и набор представлений для визуализации данных. Тезаурус опубликован в Интернете: http://www.wssanalytchem.org/ontology.

Введение

Создание онтологий - один из наиболее сложных аспектов построения семантической сети. При этом, согласно общепринятому мнению, предметные (или доменные) онтологии должны разрабатываться сообществом специалистов в той или иной области. Аналитическая химия, ставшая уже давно междисциплинарной наукой, является инструментом сертификации химического состава выпускаемой продукции, включая такую жизненно важную как продукты питания, продукты фармацевтики, оценку состояния окружающей среды и др.. В этой связи описание аналитических процедур, представление результатов анализа, а значит и соответствующая терминология строго регламентированы уполномоченными органами, как национальными (ГОСТ), так и международными (ISO, IUPAC и др.). Поэтому терминологическое обеспечение аналитической химии является заметно более продвинутым по сравнению с другими химическими научными дисциплинами.

Разработка лексической онтологии

Онтология аналитической химии в абстрактном виде должна представлять иерархию понятий аналитической химии и связей между ними. На основе многочисленных официальных документов по терминологии была разработана лексическая онтология по общим вопросам аналитической химии в виде двуязычного тезауруса ключевых терминов по общим вопросам аналитической химии, в виде электронной базы данных (БД) [1, 2]. База данных построена на реляционной модели, которая позволяет адекватно передать взаимосвязи сущностей. Главной особенностью разработанной лексической онтологии является то, что терминология из разных источников впервые интегрирована в ней на семантической основе. Это означает, что основываясь на профессиональном знании предмета аналитической химии, термины из разных источников были сгруппированы по смысловому значению (понятию), опираясь на их дефиниции из официальных документов. Понятие - это некая абстракция, смысл которой (семантика) передается в виде определенных слов (терминов) и их сочетаний на конкретном языке. В базе данных каждое понятие представлено в виде уникального числового кода - неповторяющегося целого числа. Стоит отметить, что работа по консолидации терминологии представляет собой далеко не тривиальную задачу. Она потребовала значительных затрат времени, неоднократного пересмотра ряда понятий после консультаций с другими специалистами.

Сложность работы по сведению терминологии в базу данных обусловлена также и тем, что часто при

Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2011, Воронеж, Россия, 2011.

достаточно близких дефинициях одного и того же понятия в различных источниках, соответствующие термины сильно разнятся, что приводит к путанице и разночтениям.

Для построения иерархической структуры различных источников терминологии разработано программное обеспечение для анализа загруженных данных и оценки частоты вхождений декларированных терминов в дефиниции понятий.

В настоящее время тезаурус содержит около двухсот пятидесяти понятий, более одной тысячи ста терминов, свыше пятисот дефиниций и много другой вспомогательной информации. Сюда вошли родовидовые отношения терминов (отношения «выше-ниже»), симметричная ассоциация понятий, термины разделены на ряд категорий (дескрипторы понятий, наборы синонимов различной близости к понятию, антонимы и др.). Реализация этих особенностей тезауруса заложила фундамент для разработки продвинутой онтологии на лексической основе.

Представление лексической онтологии в Интернете

Для представления лексической онтологии в Интернете развернута система сайтов на базе MS WSS 3.0. Проблема состоит в том, что MS WSS 3.0 не поддерживает реляционные связи между таблицами и межтабличные запросы. Эта система позволяет представить в Интернете лишь отдельные таблицы, при этом имеющиеся средства для представления данных ограничиваются сортировкой, выборкой и группировкой данных. Поэтому прямой перенос реляционной базы данных в систему MS WSS 3.0 хотя и возможен, но по большей части бесполезен.

Для представления данных лексической онтологии, необходимо в исходной реляционной базе данных скомпилировать сводные таблицы, содержащие всю необходимую информацию для представления данных средствами MS WSS 3.0 в том или ином ракурсе.

Для создания такой сводной таблицы, разработана система последовательных запросов к таблицам исходной БД, состоящей из семи этапов. В результате информация базы данных собирается в объемной сводной таблице. Смысл запросов состоит в том, чтобы отразить в сводной таблице все понятия аналитической химии, соответствующие дефиниции и все термины (включая их статус). Образуется разреженная таблица, с большим количеством повторяющейся информации, но зато пригодная для формирования требуемых представлений средствами MS WSS 3.0.

Далее с помощью интерактивных средств MS WSS 3.0 разработаны и размещены на дочернем сайте Научного Совета по аналитической химии (НСАХ РАН) [3] различные представления лексической онтологии. Члены совета, имеющие авторизацию на сайте, могут давать комментарии по каждой записи БД. Такой механизм обеспечит устранение возможных ошибок тезауруса и послужит

основой для развития и гармонизации терминологии по аналитической химии силами всего профессионального сообщества.

Литература

[1] Колотов В.П., Широкова В.И., Аленина М.В. Реляционная база данных как структурированное хранилище многоязычного глоссария терминов по аналитической химии. Разработка лингвистической онтологии. Труды XI Всероссийской научной конференции RCDL'2009 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (г. Петрозаводск, Россия, 17-21 сентября 2009г.). Петрозаводск: КарНЦ РАН, 2009, с.359-362.

[2] Широкова В.И., Колотов В.П., Киселева И.Н., Аленина М.В, Садовников А.А. Интегрирование терминов общего раздела аналитической химии из различных официальных документов в электронный глоссарий. Сборник трудов "Аналитическая химия - новые методы и возможности" Съезда аналитиков России и Школы молодых ученых 26-30 апреля 2010 г. Москва (пансионат "Клязьма"). М.: Издательский Дом МИСиС. 2010. С.336-339

[3] WSS- сайт по терминологии НСАХ РАН: http://www.wssanalytchem.org/ontology

Development of Lexical Ontology in Analytical Chemistry: Working out of Thesaurus as a Relational Database, Data Migration to the MS WSS 3.0 and Publication in Internet

© V.I. Shirokova, V.P. Kolotov, M.V. Alenina.

The lexical ontology in the form of the multilingual thesaurus of terms on the general issues of analytical chemistry is developed. One of the features of this thesaurus is that it is constructed applying relational model of interrelation of concepts. Terminology from different official sources is integrated on a semantic basis for the first time. The thesaurus includes sets of synonyms, different relations of terms, symmetric association of concepts.

For the thesaurus publication in the Internet by means of MS SharePoint Services (WSS 3.0) a system of queries in a relational database has been developed. The result of queries is formed over the consolidated table. A set of representations for visualization of the consolidated table data by means of MS WSS 3.0 has been developed. The thesaurus is available in the Internet: http://www.wssanalytchem.org/ontology.

♣ Работа проводится при поддержке Российского фонда фундаментальных исследований (грант N 11-03-01136-а)