

# Реляционная база данных как структурированное хранилище многоязычного глоссария терминов по аналитической химии.

## Разработка лингвистической онтологии \* ♣

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

Институт геохимии и аналитической химии им. В.И.Вернадского РАН,  
119991, ГСП-1, Москва, ул. Косыгина, 19  
[shirokova@geokhi.ru](mailto:shirokova@geokhi.ru)

### Аннотация\*

В качестве первого шага к разработке онтологии по аналитической химии создана база данных ключевых понятий в виде двуязычного русско-английского глоссария (ее расширение для других языков подразумевается). Разработана структура базы данных, проведено ее наполнение данными из различных официальных документов и выполнено ранжирование данных. Анализ результатов ранжирования (иерархии терминов) позволяет выявить противоречия и неточности терминологии и дать рекомендации по их устранению (например, необходимость уточнения дефиниций терминов). Имеется в виду, что дефиниции терминов, описывающих понятия более высокого порядка, должны включать базовые термины из глоссария, а не являться произвольным текстом, даже близким по смыслу. Каждому термину присвоен определенный статус, который обеспечивает возможность поиска понятий и терминов в случае нечетко выраженного запроса. База данных будет опубликована в Интернете на MS Windows SharePoint-сайте для ознакомления профессиональным сообществом химиков-аналитиков и в образовательных целях.

### 1 Введение

Запущен проект по созданию онтологии по аналитической химии. Работа над развернутым проектом мотивирована следующими моментами: создание предметной онтологии по аналитической

химии в связи с активно обсуждаемой концепцией семантического Интернета, подразумевающей размещение информационных ресурсов, содержащих структурированную и формализованную информацию, «понятную» компьютерам [1]; необходимость гармонизации терминологии по аналитической химии, включая электронное обеспечение научного и образовательного сообщества. Концепция семантического Интернета предполагает наличие в сети предметных онтологий (по сути словарей дефиниций, понятий и терминов той или иной области знания) и увязывание размещаемых в Интернете материалов с этими онтологиями с помощью содержательных дескрипторов XML. Онтологии, в том числе и по аналитической химии, должны создаваться сообществом специалистов.

В качестве первого шага разработана база данных ключевых понятий этой отрасли науки в виде многоязычного глоссария. Имея в виду, что методы аналитической химии часто используются для сертификации различных материалов, продуктов питания, контроля состояния окружающей среды и т.д., то описание аналитических процедур, представление результатов анализа, а значит и соответствующая терминология строго регламентируются уполномоченными органами, как национальными (ГОСТ), так и международными (ISO, IUPAC и др.). Поэтому терминологическое обеспечение аналитической химии заметно более продвинутое по сравнению с другими научными дисциплинами, где часто используется лингвистический анализ реальных текстов для выбора терминологии с последующей экспертизой для снижения информационного шума и формализации данных [3]. В аналитической химии в качестве первичных источников терминов, прошедших высококачественную профессиональную экспертизу служат утвержденные соответствующими официальными органами документы, в том числе и глоссарии терминов. Как правило, такие глоссарии

---

Труды 11<sup>й</sup> Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2009, Петрозаводск, Россия, 2009.

достаточно полны и содержат как термины, так и их синонимы (в ряде случаев антонимы), комментарии и, обязательно, дефиниции терминов.

Следует отметить, что несмотря на значительные усилия по разработке непротиворечивой системы терминологии существуют разночтения в трактовке терминов, разработанных различными организациями. Отчасти это объясняется и тем, что многие методы аналитической химии возникли на стыке с другими науками, что привело к определенному взаимовлиянию терминов, появлению терминов-синонимов, трактуемых по разному в различных областях аналитической химии. Это относится как к англоязычной, так и русскоязычной терминологии. В этой связи сведение всей официальной терминологии по аналитической химии в единый структурированный электронный источник (что является одной из целей проекта) обеспечивает удобное и мощное средство для ее анализа и, в том числе, для гармонизации терминологии в целом. Естественно, для обеспечения информационно-поисковых задач по аналитической химии глоссарий должен быть многоязычным (на текущем этапе русско-английским).

## 2 Источники информации

Нами использованы следующие официальные источники терминологии:

1. Compendium of Analytical Nomenclature: Definitive rules 1997. Orange Book. 3rd edition Inczedy, J.; Lengyel, T. and Ure, A.M. Blackwell Science, 1998 [ISBN 0-86542-6155]. On-line version: [http://www.iupac.org/publications/analytical\\_compendium](http://www.iupac.org/publications/analytical_compendium) (язык: английский),

2. International Vocabulary of Metrology – Basic and General Concepts and Associated Terms. VIM. 3rd edition. Final 2007-05-18 (язык: английский, французский),

3. ГОСТ Р 8.563-96 «Государственная система обеспечения единства измерений. Методики выполнения измерений» - М.: Издательство стандартов (язык: русский),

4. Национальный стандарт Российской Федерации. ГОСТ Р 52361-2005 «Контроль объекта аналитический. Термины и определения». М.: Стандартинформ. 2005 (язык: русский),

5. Государственный стандарт Российской Федерации. ГОСТ Р ИСО 5725.1-5725.6 «Точность (правильность и прецизионность) методов и результатов измерений». Часть 1. Основные положения и определения. М.: Издательство стандартов (язык: русский),

6. Представление результатов химического анализа. Рекомендации ИЮПАК 1994 (IUPAC. 1994. V. 66. P. 595). Перевод с англ. // Журнал аналитической химии. 1998. Т. 53. N 9. С. 999-1008 (язык: русский),

7. РМГ 61-2003 «Показатели точности,

правильности, прецизионности методик количественного химического анализа» - М., ИПК Издательство стандартов. 2004 (язык: русский).

## 3 Реляционная база данных

В результате предварительного тестирования различных способов электронной интеграции разнородных источников данных, выбор пал на использование модели реляционной базы данных как удобного средства для хранения данных, поддержания их целостности, представления отношений терминов различного ранга, обеспечения прослеживаемости истории записей, экспорта в XML-формат, публикации в Интернете и т.д. В качестве физической СУБД использован сервер MS SQL 2005.

Разработана структура базы данных, проведено ее наполнение и выполнено ранжирование.

На Рис.1 приведены основные таблицы развернутой базы данных и отношения между ними.

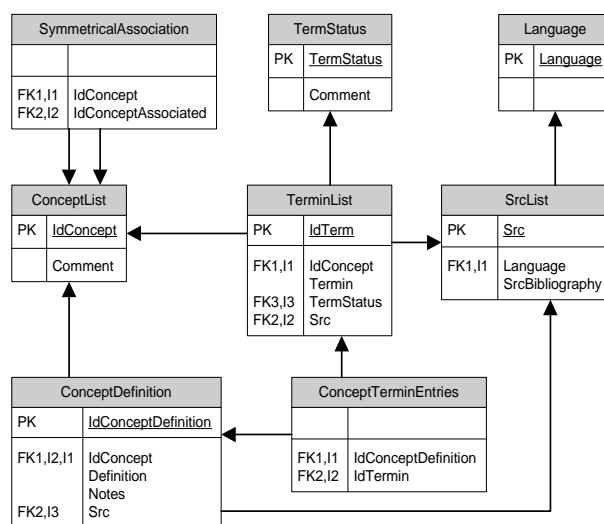


Рис.1. Структура реляционной базы данных глоссария: таблицы и поля, PK и FK(n) первичные и внешние ключи для обеспечения целостности данных, их каскадного изменения/удаления, подстановки, I(n)- индексированные поля. Стрелками показаны реляционные отношения (обычно один - ко многим).

Центральной таблицей является ConceptList (список понятий). Абстрактные понятия попросту представлены целым числом, а поле комментария позволяет дать наиболее частотное (или удобное) их словесное обозначение.

Термины и их лингвистические формы (на любом языке), соответствующие данному понятию, записаны в таблицу TerminList (фрагмент таблицы приведен на Рис. 2). Одно и то же число в поле IdConcept означает семантическую идентичность терминов.

Рис. 2. Фрагмент таблицы TerminList базы данных

IdTerm	IdConcept	Termin	Src
107	66	systematic measurement error	JCGM/WG 2
109	66	systematic error of measurement	JCGM/WG 2
110	66	systematic error	JCGM/WG 2
601	66	полная систематическая погрешность	IUPAC_1994
642	66	bias	ГОСТ Р ИСО 5725-1-2002
432	66	систематическая погрешность	ГОСТ Р ИСО 5725-1-2002
672	66	bias	IUPAC_1994
157	66	систематическая погрешность	ГОСТ Р 8.563 - 96 ГСИ

Рис. 3. Фрагмент таблицы ConceptDefinition базы данных.

IdConcept	Definition	Src
66	Component of measurement error that in replicate measurements remains constant or varies in a predictable manner	JCGM/WG 2
66	Разность между математическим ожиданием и истинным значением (со знаком)	IUPAC_1994
66	Разность между математическим ожиданием результатов измерений и истинным ( или в его отсутствие - принятым опорным ) значением	ГОСТ Р 8.563 - 96 ГСИ
66	Разность между математическим ожиданием результатов измерений и истинным (или в его отсутствие - принятым опорным) значением	ГОСТ Р ИСО 5725-1-2002

ConceptDefinition (фрагмент таблицы приведен на Рис. 3) содержит определения (дефиниции) того или иного понятия (на языке источника). Количество дефиниций определяется количеством источников данных. Кроме того, имеется ряд вспомогательных таблиц, обеспечивающих подстановку данных, например, источник терминологии и статус термина. О последней таблице есть смысл сказать отдельно. Выделены следующие особенности терминов: основной термин (предпочтительный для данного понятия); его полноценный синоним; синоним употребляемый, но устаревающий и/или не совсем точный (рекомендуется избегать использования); синоним не вошедший в нормативные документы, но, тем не менее, используемый на практике; синоним ошибочный или устаревший (должно избегать использования). Такая иерархия терминов обеспечивает возможность поиска даже нечетко выраженной информации, включая грубую оценку ее соответствия запросу.

Для того, чтобы не перегружать смысловую часть схемы (Рис.1) из нее убрана информация,

касающаяся истории редактирования данных (записи при редактировании не удаляются, имеются средства для восстановления истории модификации базы данных).

Для ранжирования данных разработано программное обеспечение (TerminRange), написанное на языке С#, включающее ряд последовательных запросов к базе данных. Результатом работы программного обеспечения являются таблица отношений терминов (ConceptListEntries), построенная на основе анализа вхождений терминов в дефиниции для определенного источника терминов. Анализ этих таблиц позволяет выявить иерархию терминов.

В базе данных в настоящее время представлены следующие отношения:

- наследуемые материнско-дочерние или родовидовые отношения. Эта наиболее многочисленная группа отношений строится автоматически с помощью программы TerminRange (табл. ConceptTerminEntries),

- симметричная ассоциация (очень близкие, но разные понятия). Эта узкая группа отношений задается вручную экспертом при компиляции базы данных, когда приходится учитывать тонкие нюансы различающие понятия (табл. SymmetricalAssociations).

Анализ иерархии терминов позволяет выявить некоторые невязки и неточности терминологии и дать рекомендации по их устранению (например, необходимость уточнения дефиниций терминов). Здесь имеется в виду, что дефиниции терминов, описывающих понятия более высокого порядка, должны включать базовые термины из глоссария, а не являться произвольным текстом, даже близким по смыслу.

База данных, включая отношения терминов, будет опубликована на MS Windows SharePoint-сайте [2] для публичного обсуждения профессиональным сообществом химиков-аналитиков.

## Литература

- [1] Berners Lee T. Semantic Web Road Map:  
<http://www.w3.org/DesignIssues/Semantic.html>
- [2] WSS- сайты Научного совета по аналитической химии Российской академии наук:  
<http://www.rusanalytchem.org/wss>
- [3] Б.В.Добров, Н.В.Лукашевич, Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Web Journal of Formal, Computational & Cognitive Linguistics - Special issue, 2006.  
[http://fccl.ksu.ru/issue\\_spec/docs/oent-kgu.doc](http://fccl.ksu.ru/issue_spec/docs/oent-kgu.doc)

## Relational database as the structured storage of a multilingual glossary of terms in analytical chemistry.

### Working out linguistic ontology

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

As the first step to working out ontology in analytical chemistry the database of key concepts in the form of a bilingual Russian-English glossary (its expansion for other languages is meant) is created. The database structure is developed, its filling by the data from various official documents is spent and ranging of the data is executed. The analysis of results of ranging (hierarchy of terms) allows to reveal contradictions and discrepancies of terminology and to make recommendations about their elimination (for example, necessity of specification of definitions of terms). Means that definitions of the terms describing concepts of higher order, should include base terms from a glossary, instead of be any text, even faithful. To each term the certain status which provides possibility of search of concepts and terms in case of indistinctly expressed inquiry is appropriated. The database will be published on the Internet on MS Windows a SharePoint-site for acquaintance by professional community of chemists-analysts and in the educational purposes.

---

Работа проводится при поддержке Российского фонда фундаментальных исследований (грант N 08-03-00893)