Статьи в формате Enhanced HTML на сайте Royal Society of Chemistry

В настоящее время компьютер воспринимает текст формально, как последовательность букв, цифр, математических и иных знаков. В частности, современные поисковые машины извлекают те документы, в которых обнаруживается набор символов, заданный в запросе.

Научить компьютер распознавать смысл написанного (**семантику**) и на этой основе создать веб нового поколения, **семантический веб**,— над такой проблемой работают многочисленные исследователи, приверженцы новых идей создателя *WWW* Тима Бернерса-Ли.

Разные методы апробируются для решения проблемы; один из них — обогащение текста документа **метаданными** (данными о данных). Метаданные предназначены компьютеру и содержат в себе пояснения смысла используемых в тексте терминов.

Каждая предметная область имеет свою структуру знаний — **онтологию**. Онтология включает в себя набор понятий и взаимосвязи этих понятий (*упрощенный пример фрагмента онтологии неорганической химии*: химический элемент ↔ *d*-элемент I группы ↔ серебро ↔ хлорид серебра). Указывая в метаданных место данного термина в соответствующей онтологии, мы сообщим компьютеру сведения и о смысловом значении этого слова, и о смысловых связях данного текста с другими документами.

Пример. С помощью метаданных можно показать, что термины *срэбра*, *серебро* и *silver* означают одно и то же, а слово *золото* в сочетаниях «*металл золото*», «*черное золото*», «*молчание — золото*» имеет разный смысл (как термин разных онтологий); что между понятиями *золото* и *серебро* существует больше общего, чем между близкими по написанию *золото* и *болото*.

Практическая реализация идей семантического веба — дело будущего, тем не менее, подготовительные работы ведутся уже теперь. В области химии наиболее целостными выглядят результаты исполнения проекта **Project Prospect** Королевского химического общества (*RSC*).

RSC Project Prospect

Для того, чтобы компьютер мог распознавать научные компоненты в статьях, опубликованных в журналах *RSC*, статьи обогащают метаданными. Технические редакторы анализируют текст, находят в нем слова и комбинации слов, несущие смысловую научную нагрузку (названия веществ, научные понятия и др.), прикрепляют к этим словам комментарии и добавляют гиперсвязи, ведущие к дополнительным электронным ресурсам, например, базам данных. В ходе семантического обогащения учитываются химическая и биологическая онтологии.

Результаты работы публикуются на сайте *RSC* в формате, который здесь называется **Enhanced HTML**. В оглавлениях журналов статьи, содержащие материал с метаданными, имеют пометку *RSC Prospect*:

Дополнительная информация, включаемая редакторами в *Enhanced HTML*, предназначается не только компьютеру, но и человеку, что делает эту версию статьи привлекательной для читателя.

Ранее файлы *Enhanced HTML* были доступны только подписчикам. Теперь у нас появилась возможность детально ознакомиться с методикой, которую использует *Royal Society of Chemistry* для семантического обогащения публикаций.

В 2009 г. году каждый новый выпуск журнала **Metallomics** (статьи в форматах *PDF*, *HTML* и ***Enhanced HTML***) будет открыт для неподписчиков в течение всего периода до выхода очередного номера.

Metallomics

http://www.rsc.org/Publishing/Journals/mt/index.asp

Оглавление свежего выпуска размещено на Главной странице журнала. От названия статьи или ее кода *DOI* читатель переходит на страницу, содержащую реферат и ссылки на полные материалы статьи в форматах *PDF*, *HTML* и *Enhanced HTML*.

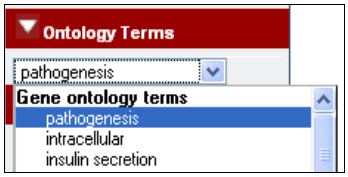
Enhanced HTML

При выводе на экран обогащенной *HTML*-версии статьи читатель видит в правом верхнем углу полупрозрачное навигационное окошко **Toolbox**, содержимое которого проявляется при наведении курсора.

Блок Tools and Resources стандартен для каждой из *HTML*-версий, остальная часть навигационных инструментов специфична и имеется только в *Enhanced HTML*.

Меню Navigation предназначено для быстрого перехода к соответствующему разделу статьи.

Меню Ontology Terms содержит перечень терминов, встречающихся в данной статье и принадлежащих нескольким биологическим и одной химической онтологиям.

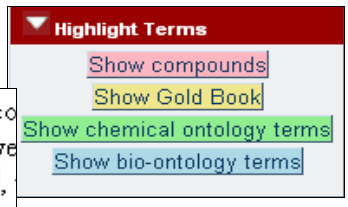


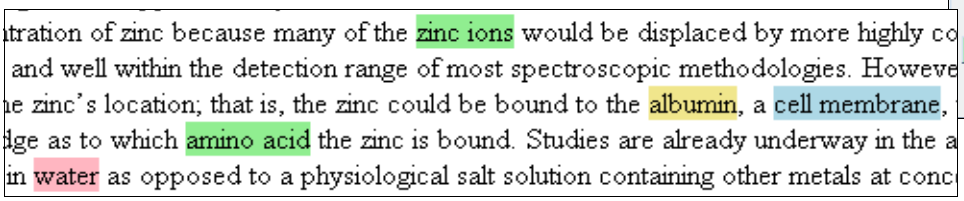
При щелчке по термину открывается окно, в котором содержится онтологическая и поясняющая информация: определение термина, идентификационные коды, синонимы названия, гиперссылки на статьи в журналах *RSC*, где упоминается этот термин, гиперссылка на соответствующую запись в базе данных, описывающей данную онтологию.

Используемая здесь химическая онтология называется **ChEBI** (*Chemical Entities of Biological Interest*, т. е. *Химические объекты, интересные для биологии*). Под объектами понимаются, как правило, низкомолекулярные вещества, отдельные атомы, молекулы, ионы, ионные пары, радикалы, комплексные частицы и т. п.

ChEBI — это одновременно и словарь химических объектов, и классификационная схема.

Меню Highlight Terms позволяет выделить цветом в тексте статьи термины четырех типов.





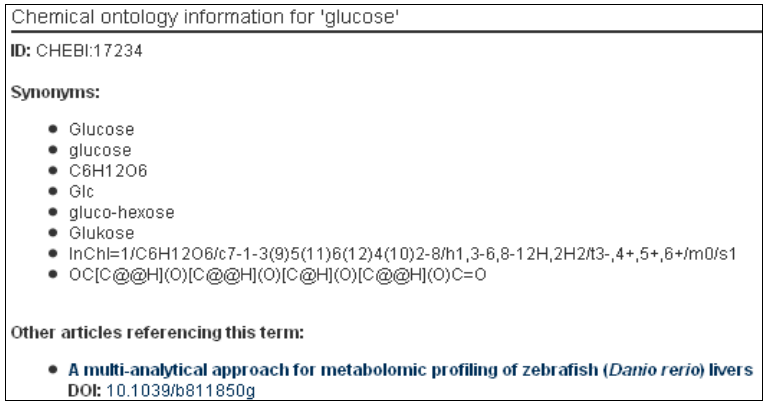
Кнопка Show compounds выделяет в тексте розовым фоном названия химических веществ. Если щелкнуть по такому названию, в новом окне приводится следующая информация о веществе:

- определение термина;
- синонимы названия;
- коды *SMILES*, *InChI*, *InChI Key*;
- ссылка на файл в формате *CML* (структура объекта);
- двумерная графическая формула;
- гиперсвязи к статьям проекта *Project*, в которых упоминается это вещество,
- ссылка на информацию об этом веществе в справочной базе данных *PubChem*,
- ссылка на информацию о веществе в патентной базе данных *SureChem*.

Кнопка Show Gold Book выделяет желтым цветом термины, упоминаемые в *Compendium of Chemical Terminology* — справочнике ИЮПАК по химической терминологии. При щелчке по слову соответствующий фрагмент справочника открывается в отдельном окне браузера.

Кнопка Show chemical ontology terms зеленым цветом выделяет термины, упоминаемые в химической онтологии *ChEBI*. При щелчке по такому термину открывается новое окно, содержащее следующие сведения:

- определение термина;
- код объекта в *ChEBI*;
- синонимы названия;
- гиперсвязи к другим статьям;
- ссылка на информацию об этом объекте в *ChEBI*.


Chemical ontology information for 'glucose'

ID: CHEBI:17234

Synonyms:

- Glucose
- glucose
- C6H12O6
- Glc
- gluco-hexose
- Glukose
- InChI=1/C6H12O6/c7-1-3(9)5(11)6(12)4(10)2-8/h1,3-6,8-12H,2H2/t3-,4+,5+,6+/m0/s1
- OC[C@@H](O)[C@@H](O)[C@H](O)[C@@H](O)C=O

Other articles referencing this term:

- **A multi-analytical approach for metabolomic profiling of zebrafish (*Danio rerio*) livers**
  DOI: 10.1039/b811850g

Кнопка Show bio-ontology terms голубым цветом выделяет биомедицинские термины, имеющиеся в онтологиях генов, клеток и аминокислотных рядов.

Если кнопки меню Highlight Terms не нажаты, упомянутые выше термины подсвечиваются в тексте при наведении на них курсором.

В **списке литературы** ссылки External Links ведут к онлайновым публикациям (непосредственно либо через *CrossRef*) или к рефератам в реферативной базе данных *ChemPort*.

RSC Prospect Structure Search

В настоящее время проходит бета-тестирование поисковой программы, использующей структурную формулу в качестве поискового термина.

Поиск ведется только по статьям, содержащим метаданные.

Ссылка на соответствующий поисковый бланк (RSC Prospect Structure Search) находится в навигационном меню всех страниц раздела *Journals* сайта *RSC* (в левой колонке).

В качестве поискового термина может быть:

- строка *SMILES*;
- формула, сформированная в *ChemDraw*, *ISIS Draw*;
- формула, сформированная с помощью встроенного апплета *MarvinSketch*.

В результатах поиска приводится:

- название вещества, его синонимы;
- код *InChI*;
- двумерная графическая формула;
- ссылки на статьи (только из базы *Project Prospect*), в которых упоминается это вещество;
- ссылка Show close matches на статьи, частично соответствующие запросу.

В тексте упоминались следующие сайты:

Royal Society of Chemistry http://www.rsc.org/
RSC Project Prospect http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp
Metallomics http://www.rsc.org/Publishing/Journals/mt/index.asp
RSC Prospect Structure Search http://www.rsc.org/Publishing/Journals/structuresearch.asp
IUPAC Gold Book http://goldbook.iupac.org/index.html
Chemical Entities of Biological Interest (ChEBI) http://www.ebi.ac.uk/chebi/init.do
ChemPort http://chemport.cas.org/
CrossRef http://www.crossref.org/
PubMed http://www.ncbi.nlm.nih.gov/pubmed/
SureChem http://www.surechem.org/