НАСТРАИВАЕМАЯ СИСТЕМА ЭКСПОРТА ДАННЫХ МОЛЕКУЛЯРНОЙ СПЕКТРОСКОПИИ В XML

А.В. Козодоев
Институт оптики атмосферы СО РАН, Томск
e-mail: kav@iao.ru

АННОТАЦИЯ

Представлен обзор системы экспорта данных в формат XML в распределенной информационно-вычислительной системе «Молекулярная спектроскопия».

Молекулярная спектроскопия широко используется в прикладных исследованиях во многих областях физики. Предметом изучения молекулярной спектроскопии являются спектральные свойства молекул. В настоящее время хорошо изучены всего несколько десятков молекул, а детальное изучение свойств молекул не закончено до сих пор. Связано это с тем обстоятельством, что спектральные свойства молекул описываются большим объёмом данных, только описание спектральных линий может иметь сотни тысяч линий, каждая из которых описывается десятком параметров. Для эффективной работы с такими объёмными данными необходима их систематизация, имея которую можно создать соответствующие программные средства для обработки данных. Систематизация спектральных данных проведена теоретиками несколько десятков лет назад. Данные характеризующие спектральные свойства молекул собираются в банках данных, таких как HITRAN [1] и GEISA [2]. Наполнение банков данных продолжается, и этот процесс, так же как и изучение спектральных свойств молекул, далёк от завершения. Наполнению банков данных способствует и современное состояние вычислительной техники — растёт количество расчетных спектров и объём получаемых данных. Растет также число исследовательских групп. Все это указывает на необходимость использования современных информационных технологий для коллективной работы с информацией.

Начиная с 80-х годов в ИОА СО РАН ведутся разработки информационных систем в области молекулярной спектроскопии [3]. С появлением персональных компьютеров был пройден путь от систем работающих на клиентском месте (например, [4]) до доступных в Интернет информационных ресурсов (http://spectra.iao.ru) [5], опирающегося на известные банки спектроскопических данных HITRAN и GEISA, и распределённой информационно-вычислительной системы «Молекулярная спектроскопия» (http://www.saga.iao.ru/).

Хранение данных в ИВС «Молекулярная спектроскопия» осуществляется с помощью реляционной СУБД, что хорошо подходит для работы с данными в пределах ИВС. Однако, для передачи данных между различными ИВС или выгрузке данных пользователю можно использовать множество вариантов записи данных в файле. Это могут быть различные как двоичные так и текстовые форматы. Для работы с данными представленными в текстовом формате, пользователю, кроме самих данных, необходимо предоставить как описание их структуры, так и минимальное описание самих данных. Файлы, удовлетворяющие этим требованиям, можно создавать на основе стандарта XML[7].

Файл созданный на основе стандарта XML может содержать как данные, так и их описание. Описание данных может быть произведено как с помощью тегов имеющих

смысловое значение, что позволяет человеку, глядя на XML файл, понять какие данные там представлены, так и с помощью текстовых аннотаций включённых в документ в виде комментариев или специальных записей содержащих информацию для человека и игнорируемых при машинной обработке.

Создание XML файла для конкретной структуры данных и конкретного формата XML трудности не составляет. Для этого необходимо заполнять заранее заданный шаблон XML документа данными. Обычно это делается с помощью специально созданной подпрограммы, предназначенной только для этой цели. Но в таком случае любое изменение структуры XML документа приводит к необходимости изменения кода подпрограммы.

ИВС «Молекулярная спектроскопия» хранит несколько типов данных предметной области, различающихся структурой, количеством величин, перечнем обязательных и не обязательных величин, и каждый из этих типов имеет свой формат представления в XML. Таким образом, в зависимости от выбранных для экспорта из ИВС данных, необходимо генерировать XML заданного вида и содержащий выбранные данные. Так же желательно иметь возможность изменения или добавления новых XML форматов для экспорта без изменения подпрограммы экспорта. Последнее требование является существенным, так как из него вытекает, что система экспорта должна быть настраиваемой без участия программиста.

Структура и свойства данных предметной области ограничивают возможные варианты XML форматов для экспорта. Эти ограничения приводят к следующим свойствам, которыми должен обладать XML документ. Документ может иметь часть содержащую характеристики данных, и часть содержащую данные. Данные сгруппированы блоками в соответствии с типом данных предметной области. Например, величины характеризующие конкретную спектральную линию находятся в одном блоке.

Учитывая вышеприведённые требования и ограничения был разработан алгоритм генерации XML документов на основе шаблонов. Рассмотрим работу алгоритма на примере характеристик спектрального перехода.

Далее представлен вариант конечного XML документа содержащий частоту перехода, ошибку частоты и квантовые числа идентифицирующие один спектральный переход.

```
<?xml version="1.0" encoding="UTF-8"?>
<T6 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://wadis.saga.iao.ru/data/xsd/tasks/version3/task/T6.xsd">
<_l>
  <VWN error="5000.0e-6">10787.327</VWN>
  <Transition>
    <TrNM>
      <Upper>
        <v1>2</v1><v2>1</v2><v3>2</v3><J>14</J><Ka>2</Ka><Kc>13</Kc>
      </Upper>
      <Lower>
        <v1>0</v1><v2>0</v2><v3>0</v3><J>15</J><Ka>1</Ka><Kc>14</Kc>
      </Lower>
    </TrNM>
  </Transition>
</_l>
</T6>
```

Для создания приведенного выше документа использованы шаблоны двух типов — шаблон документа и шаблон данных. Шаблон документа включает в себя заголовок XML документа и корневой тег (см. пример далее). Шаблоны данных включают только теги

относящиеся к блоку данных (тег <_l>) и конкретной величине. Место в шаблоне для подстановки значения величины указывается комбинацией символов «%s». В данном примере шаблоны получаются следующие.

Шаблон документа:

```
<?xml version="1.0" encoding="UTF-8"?>
<T6 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://wadis.saga.iao.ru/data/xsd/tasks/version3/task/T6.xsd">
%s
</T6>
```

Шаблон частоты перехода:

```
<_l>
  <VWN >%s</VWN>
</_l>
```

Шаблон ошибки частоты перехода:

```
<_l>
  <VWN error="%s"></VWN>
</_l>
```

Шаблон верхнего квантового числа «v1»:

```
<_l>
  <Transition>
    <TrNM>
      <Upper>
        <v1>%s</v1>
      </Upper>
    </TrNM>
  </Transition>
</_l>
```

Шаблоны остальных величин получаются аналогично.

Генерация XML документа на основе таких шаблонов происходит следующим образом. Сначала генерируются блоки данных — шаблоны величин заполняются соответствующими данными, затем производится рекурсивное слияние шаблонов основываясь на именах тегов и атрибутов. Таким образом получаются блоки данных. Затем блоки данных вставляются в шаблон документа. В результате чего получается XML документ с требующимися пользователю данными и заданного формата.

На данный момент шаблон может быть создан или изменён только пользователем системы с привилегиями администратора. Использовать шаблоны может любой пользователь системы.

С точки зрения программной реализации основную сложность представляет рекурсивное слияние шаблонов. Были опробованы два варианта реализации на языке PHP: алгоритм работающий с текстовым представлением и алгоритм использующий библиотеку для работы с DOM (Document Object Model). Использование библиотеки DOM позволило избавиться от множества операций со строками, и использовать некоторые оптимизации что привело к значительному увеличению производительности. Обработка 20 тыс. блоков данных, соответствующих параметрам спектральных линий, алгоритмом с использованием строк выполнялась около 10 минут. Реализация с использованием DOM обрабатывала тот же объём данных примерно за 5-10 секунд.

В заключении можно сказать, что разработанное программное обеспечение на основе вышеприведённого алгоритма позволяет задавать шаблоны для XML документов и генерировать XML документ на их основе. Это позволяет производить экспорт данных представленных в РИВС «Молекулярная спектроскопия» в файлы формата XML.

ЛИТЕРАТУРА

1. HITRAN, http://www.hitran.com.
2. *Jacquinet-Husson N., Arie E.*, et al, The 1997 spectroscopic GEISA databank, JQSRT, v.62, pp. 205-254, 1999. (http://www.ara.polytechnique.fr).
3. *Войцеховская О.К., Макушкин Ю.С., Попков А.И., Розина А.В., Руденко В.П., Трифонова Н.Н., Яковлев Н.Е.* Структура и принципы реализации подсистемы формирования банка параметров спектральных данных. // Тезисы докладов 6 Всесоюзного симпозиума по молекулярной спектроскопии высокого и сверхвысокого разрешения, Томск, 1982, ч.2., с. 42-44.
4. *Golovko V.F., Nikitin A.V., Chursin A.A., Tyuterev Vl.G.*, Information system AIRSENTRY for modeling atmospheric IR-spectra and radiation transmission in the atmosphere, ADBIS'95 Proc. The 2-nd Int. Workshop, v.2, Moscow, 1995, p.12-14.
5. *Бабиков Ю.Л., Барб А., Головко В.Ф., Тютерев Вл.Г.*, Интернет-коллекции по молекулярной спектроскопии, Сборник трудов 3 Всероссийской конференции по электронным библиотекам, Петрозаводск, 2001, с.183-187.
6. *Козодоев А.В.*, Система загрузки данных в распределённой информационной системе "Молекулярная спектроскопия", Материалы IV Всероссийской конференции молодых учёных «Материаловедение, технологии и экология в 3-м тысячелетии».- Томск: Изд. Института оптики атмосферы СО РАН, 2009.-656с.,
7. Extensible Markup Language (XML), http://www.w3.org/TR/xml/