New Tools for Storing and Accessing Spectroscopic Data

The Development of an XML Schema for the HITRAN Database

Dr Christian Hill

Department of Physics and Astronomy,

UCL

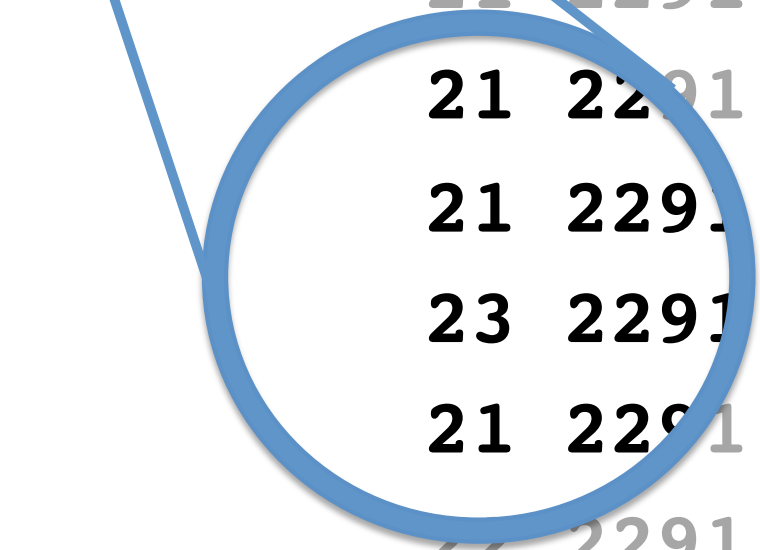
HITRAN format since 2004

```
...
 21 2291.946330 4.940E-29 3.414E-02.06820.086 3595.59060.78-.007000       4 0 0 02       2 1 1 03                Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.947360 1.660E-28 3.650E+02.06980.093 5204.94610.780.000400       0 4 4 21       0 4 4 11                R 24f     4455501221 1 1 1 4    51.0   49.0
 21 2291.953028 4.614E-27 2.522E+00.06610.073 3633.90910.760.000000       0 4 4 11       0 4 4 01                Q 49f     454550 2 2 1 1 1 0*   99.0   99.0
 23 2291.957370 6.450E-27 1.504E+02.08930.123 2501.49690.750.000000       2 0 0 13       2 0 0 03                R  1e     455550 2 2 1 1 1 0     5.0    3.0
 21 2291.959750 4.980E-28 4.906E-01.06920.091 3634.14400.78-.007000       2 1 1 12       2 1 1 02                Q 27e     4455501221 1 1 1 4    55.0   55.0
 22 2291.967650 4.920E-28 1.880E+02.06030.060 4140.44010.66-.007100       1 1 1 11       1 1 1 01                R 73e     4455501221 1 1 1 4   298.0  294.0
 21 2291.974450 3.600E-29 1.790E+02.06660.075 5496.45810.77-.004490       0 7 7 11       0 7 7 01                R 45e     4455501221 1 1 1 4    93.0   91.0
 21 2291.977720 5.480E-28 8.715E-04.06820.086 2345.92090.78-.007000       3 0 0 01       1 1 1 02                Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.986760 1.620E-27 1.084E+00.07280.100 3473.12120.74-.007000       2 1 1 12       2 1 1 02                Q 18f     4455501221 1 1 1 4    37.0   37.0
 27 2291.987280 4.040E-30 1.946E+02.06640.074 3386.79000.77-.004430       0 4 4 11       0 4 4 01                R 46e     4355501221 1 1 1 4    95.0   93.0
...
```

- ASCII text format: one line of 160 bytes per transition;
- Fixed-width formats for data fields: Fortran-friendly;
- Total database size (without supplementary data): 440 MB.

HITRAN format since 2004

```
...
 21 2291.946330 4.940E-29 3.414E-02.06820.086 3595.59060.78-.007000      4 0 0 02      2 1 1 03              Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.947360 1.660E-28 3.650E+02.06980.093 5204.94610.780.000400      0 4 4 21      0 4 4 11              R 24f     4455501221 1 1 1 4    51.0   49.0
 21 2291.953028 4.614E-27 2.522E+00.06610.073 3633.90910.760.000000      0 4 4 11      0 4 4 01              Q 49f     454550 2 2 1 1 1 0*   99.0   99.0
 23 2291.957170 6.450E-27 1.504E+02.08930.123 2501.49690.750.000000      2 0 0 13      2 0 0 03              R  1e     455550 2 2 1 1 1 0     5.0    3.0
 21 2291.959750 1.980E-28 4.906E-01.06920.091 3634.14400.78-.007000      2 1 1 12      2 1 1 02              Q 27e     4455501221 1 1 1 4    55.0   55.0
 22 2291.967650 4.920E-28 1.880E+02.06030.060 4140.44010.66-.007100      1 1 1 11      1 1 1 01              R 73e     4455501221 1 1 1 4   298.0  294.0
 21 2291.974450 3.600E-29 1.790E+02.06660.075 5496.45810.77-.004490      0 7 7 11      0 7 7 01              R 45e     4455501221 1 1 1 4    93.0   91.0
 21 2291.977720 5.480E-28 8.715E-04.06820.086 2345.92090.78-.007000      3 0 0 01      1 1 1 02              Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.986760 1.620E-27 1.084E+00.07280.100 3473.12120.74-.007000      2 1 1 12      2 1 1 02              Q 18f     4455501221 1 1 1 4    37.0   37.0
 27 2291.987280 4.040E-30 1.946E+02.06640.074 3386.79000.77-.004430      0 4 4 11      0 4 4 01              R 46e     4355501221 1 1 1 4    95.0   93.0
...
```



Molecule ID

Isotopologue ID

HITRAN format since 2004

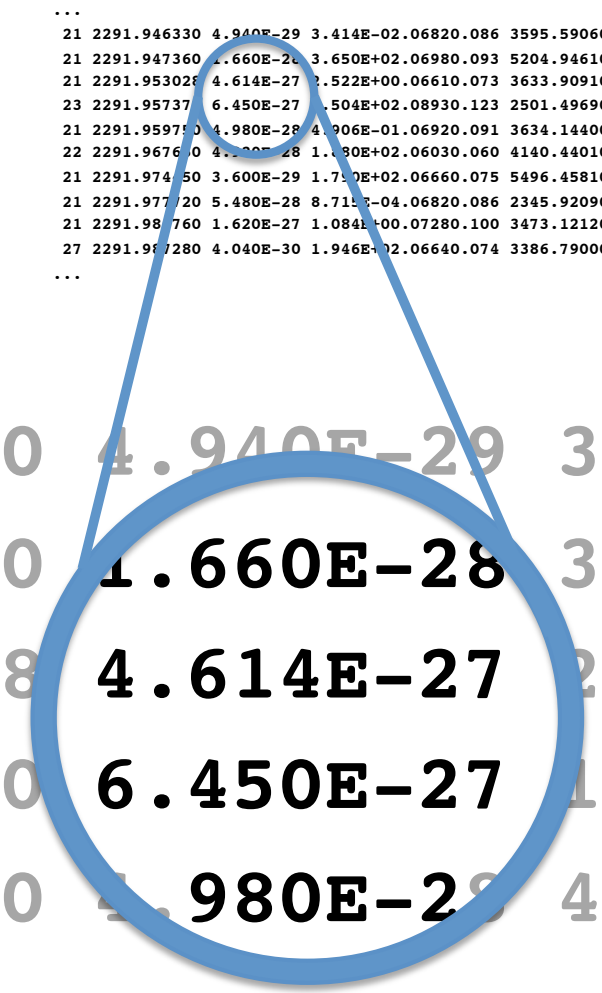
```
...
21 2291.946330 1.940E-29 3.414E-02.06820.086 3595.59060.78-.007000      4 0 0 02      2 1 1 03              Q 32f     4455501221 1 1 1 4    65.0   65.0
2  2291.947360 1.   E-28 3.650E+02.06980.093 5204.94610.780.000400      0 4 4 21      0 4 4 11              R 24f     4455501221 1 1 1 4    51.0   49.0
2  2291.953028 4.61   -27 2.522E+00.06610.073 3633.90910.760.000000      0 4 4 11      0 4 4 01              Q 49f     454550 2 2 1 1 1 0*   99.0   99.0
2  2291.957370 6.450E-   1.504E+02.08930.123 2501.49690.750.000000      2 0 0 13      2 0 0 03              R  1e     455550 2 2 1 1 1 0     5.0    3.0
2  2291.959750 4.980E-28 4.906E-01.06920.091 3634.14400.78-.007000      2 1 1 12      2 1 1 02              Q 27e     4455501221 1 1 1 4    55.0   55.0
2  229   .967650 4.920E-28 1.880E+02.06030.060 4140.44010.66-.007100      1 1 1 11      1 1 1 01              R 73e     4455501221 1 1 1 4   298.0  294.0
2  2291.974450 3.600E-29 1.790E+02.06660.075 5496.45810.77-.004490      0 7 7 11      0 7 7 01              R 45e     4455501221 1 1 1 4    93.0   91.0
2  2291.977720 5.480E-28 8.715E-04.06820.086 2345.92090.78-.007000      3 0 0 01      1 1 1 02              Q 32f     4455501221 1 1 1 4    65.0   65.0
2  2291.986760 1.620E-27 1.084E+00.07280.100 3473.12120.74-.007000      2 1 1 12      2 1 1 02              Q 18f     4455501221 1 1 1 4    37.0   37.0
2  2291.987280 4.040E-30 1.946E+02.0664 .074 3386.79000.77-.004430      0 4 4 11      0 4 4 01              R 46e     4355501221 1 1 1 4    95.0   93.0
...
```


21 2291.946330 4.940E-29 3.414E-02
2291.947360
2291.953028
2291.957370
2291.959750
22 2291.967650 4.920E-28 1.880E+02

Transition Frequency, ν /cm^{-1}

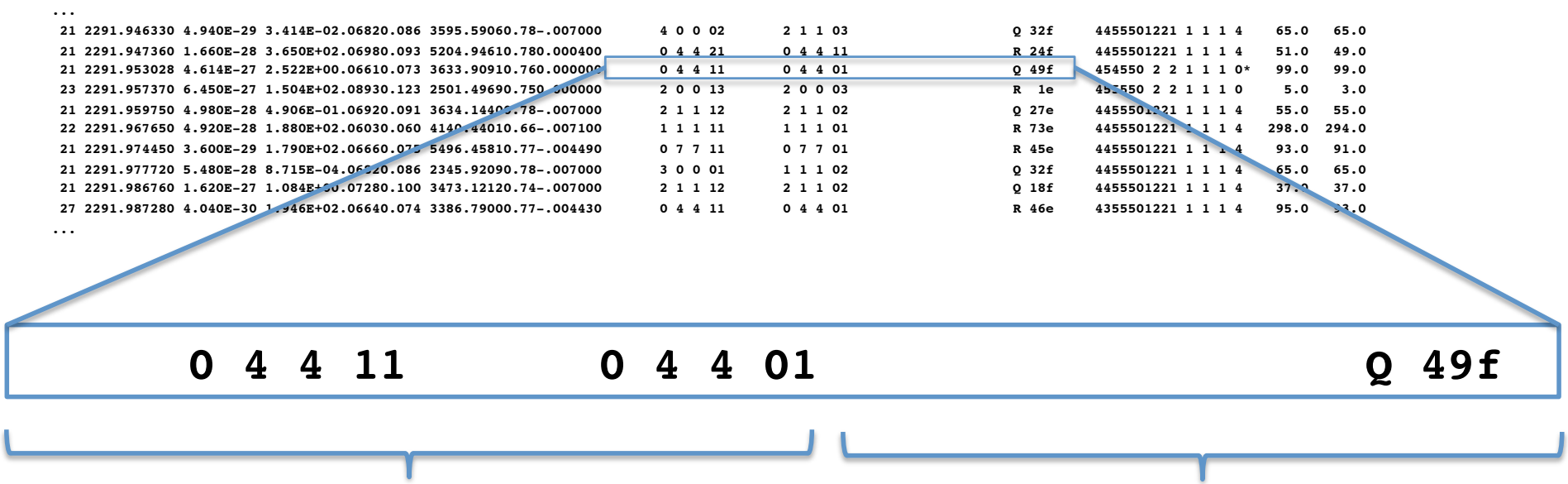
HITRAN format since 2004

```
...
 21 2291.946330 4.940E-29 3.414E-02.06820.086 3595.59060.78-.007000       4 0 0 02       2 1 1 03                Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.947360 1.660E-28 3.650E+02.06980.093 5204.94610.780.000400       0 4 4 21       0 4 4 11                R 24f     4455501221 1 1 1 4    51.0   49.0
 21 2291.95302  4.614E-27 2.522E+00.06610.073 3633.90910.760.000000       0 4 4 11       0 4 4 01                Q 49f     454550 2 2 1 1 1 0*   99.0   99.0
 23 2291.95737  6.450E-27 1.504E+02.08930.123 2501.49690.750.000000       2 0 0 13       2 0 0 03                R  1e     455550 2 2 1 1 1 0     5.0    3.0
 21 2291.9597   4.980E-28 4.906E-01.06920.091 3634.14400.78-.007000       2 1 1 12       2 1 1 02                Q 27e     4455501221 1 1 1 4    55.0   55.0
 22 2291.9676   4.    -28 1.   0E+02.06030.060 4140.44010.66-.007100       1 1 1 11       1 1 1 01                R 73e     4455501221 1 1 1 4   298.0  294.0
 21 2291.974    3.600E-29 1.7  0E+02.06660.075 5496.45810.77-.004490       0 7 7 11       0 7 7 01                R 45e     4455501221 1 1 1 4    93.0   91.0
 21 2291.977    5.480E-28 8.71 E-04.06820.086 2345.92090.78-.007000       3 0 0 01       1 1 1 02                Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.98     1.620E-27 1.084E+00.07280.100 3473.12120.74-.007000       2 1 1 12       2 1 1 02                Q 18f     4455501221 1 1 1 4    37.0   37.0
 27 2291.9      4.040E-30 1.946E+02.06640.074 3386.79000.77-.004430       0 4 4 11       0 4 4 01                R 46e     4355501221 1 1 1 4    95.0   93.0
...
```



Transition Strength,
S /cm^{-1}(molec.cm^{-2})

HITRAN format since 2004

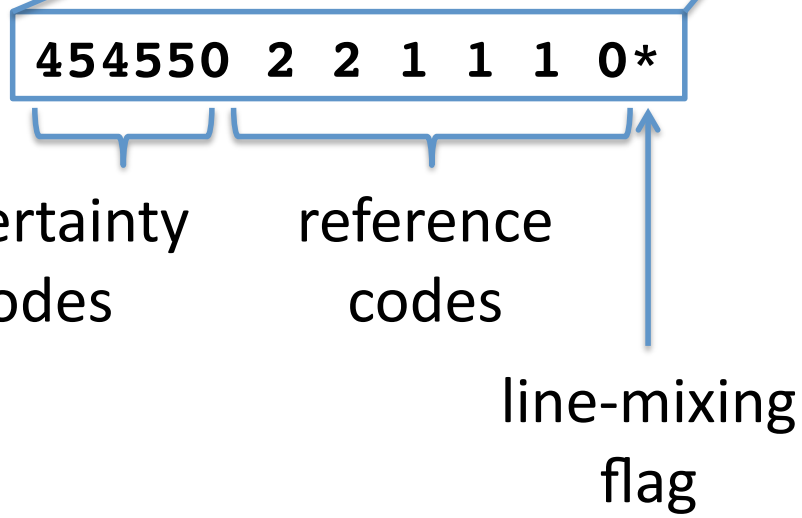


"global" quanta:
vibrational / electronic

"local" quanta:
rotational, symmetry

HITRAN format since 2004

```
...
 21 2291.946330 4.940E-29 3.414E-02.06820.086 3595.59060.78-.007000      4 0 0 02      2 1 1 03              Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.947360 1.660E-28 3.650E+02.06980.093 5204.94610.780.000400      0 4 4 21      0 4 4 11              R 24f     4455501221 1 1 1 4    51.0   49.0
 21 2291.953028 4.614E-27 2.522E+00.06610.073 3633.90910.760.000000      0 4 4 11      0 4 4 01              Q 49f     454550 2 2 1 1 1 0*   99.0   99.0
 23 2291.957370 6.450E-27 1.504E+02.08930.123 2501.49690.750.000000      2 0 0 13      2 0 0 03              R 1e      455550 2 2 1 1 1 0    5.0    3.0
 21 2291.959750 4.980E-28 4.906E-01.06920.091 3634.14400.78-.007000      2 1 1 12      2 1 1 02              Q 27e     4455501221 1 1 1 4    55.0   55.0
 22 2291.967650 4.920E-28 1.880E+02.06030.060 4140.44010.66-.007100      1 1 1 11      1 1 1 01              R 73e     4455501221 1 1 1 4    298.0  294.0
 21 2291.974450 3.600E-29 1.790E+02.06660.075 5496.45810.77-.004490      0 7 7 11      0 7 7 01              R 45e     4455501221 1 1 1 4    93.0   91.0
 21 2291.977720 5.480E-28 8.715E-04.06820.086 2345.92090.78-.007000      3 0 0 01      1 1 1 02              Q 32f     4455501221 1 1 1 4    65.0   65.0
 21 2291.986760 1.620E-27 1.084E+00.07280.100 3473.12120.74-.007000      2 1 1 12      2 1 1 02              Q 18f     4455501221 1 1 1 4    37.0   37.0
 27 2291.987280 4.040E-30 1.946E+02.06640.074 3386.79000.77-.004430      0 4 4 11      0 4 4 01              R 46e     4355501221 1 1 1 4    95.0   93.0
...
```



HITRAN format since 2004

Limitations:

- Hard to extend to include *e.g.*
  - quantum numbers for complex states,
  - line-mixing data,
  - new line-broadening species (*e.g.* H_2O, H_2),
  - parameters for lineshapes other than Voigt;
- Many states duplicated (participate in more than one transition);
- Arbitrary default entries indicating unavailable data (*e.g.* -1. for lower-state energy);
- Errors and inconsistencies hard to identify (format contains no *semantic* information).

VAMDC

- **V**irtual **A**tomic and **M**olecular **D**ata **C**entre;
- EU Project funded under Framework Programme 7: *Research Infrastructure*;
- Aims to build “an interoperable e-infrastructure for the exchange of atomic and molecular data”;
- Development of tools for storing, searching and manipulating AM data from many different sources.

Relational Database Model

States Table

StateID	Energy	Uncertainty	J	Ka	Kc	v1	v2	v3	…
S1-H2O-1	1394.8142	0.0001	7	7	1	0	0	0	…
S2-H2O-1	1590.69006	0.00001	8	7	2	0	0	0	…
S3-H2O-1	2433.8003	0.0001	12	6	7	0	0	0	…
S4-H2O-1	3441.03948	0.00001	11	3	8	0	1	0	…
…									

Transitions Table

TransID	UpperStateID	LowerStateID	ν	S	…
L1-H2O-1	S2-H2O-1	S1-H2O-1	195.87586	2.743E-25	…
L2-H2O-1	S3-H2O-1	S7-H2O-1	273.38293	1.324E-24	…
L3-H2O-1	S4-H2O-1	S12-H2O-1	298.01003	8.321E-25	…
L4-H2O-1	S9-H2O-1	S29-H2O-1	312.96733	5.565E-25	…
…					

Relational Database Model

States Table

StateID	Energy	Uncertainty	*J*	*Ka*	*Kc*	*v1*	*v2*	*v3*	…
S1-H2O-1	1394.8142	0.0001	7	7	1	0	0	0	…
S2-H2O-1	1590.69006	0.00001	8	7	2	0	0	0	…
S3-H2O-1	2433.8003	0.0001	12	6	7	0	0	0	…
S4-H2O-1	3441.03948	0.00001	11	3	8	0	1	0	…
…									

Transitions Table

TransID	UpperStateID	LowerStateID	ν	S	…
L1-H2O-1	S2-H2O-1	S1-H2O-1	195.87586	2.743E-25	…
L2-H2O-1	S3-H2O-1	S7-H2O-1	273.38293	1.324E-24	…
L3-H2O-1	S4-H2O-1	S12-H2O-1	298.01003	8.321E-25	…
L4-H2O-1	S9-H2O-1	S29-H2O-1	312.96733	5.565E-25	…
…					

Relational Database Model

- Based on MySQL (free, open-source)
- Query using SQL = Structured Query Language
- Web interface:
  - http://msslsc.mssl.ucl.ac.uk/
- Output formats:
  - Original HITRAN format (.par)
  - ASCII-text table of tab-delimited columns (.txt)
  - XSAMS (.xml) …

XSAMS

- Under development by the IAEA
- An XML format for distributing Atomic and Molecular Spectroscopic Data
- Enforces good practice:
  - Data sources (*e.g.* literature references)
  - Uncertainties
  - Compulsory units

XSAMS – Example: a molecular state of H2O

```
<MolecularState stateID="S14-H2O-1">
   <Description>A state of H2(16O)</Description>
   <MolecularStateCharacterisation>
      <StateEnergy>
         <Value units="1/cm”>1813.223400</Value>
      </StateEnergy>
      <TotalStatisticalWeight>69</TotalStatisticalWeight>
   </MolecularStateCharacterisation>
   <nltcs:QNs>
      <nltcs:ElecStateLabel>X</nltcs:ElecStateLabel>
      <nltcs:J>11</nltcs:J>
      <nltcs:Ka>3</nltcs:Ka>
      <nltcs:Kc>8</nltcs:Kc>
      <nltcs:v1>0</nltcs:v1>
      <nltcs:v2>0</nltcs:v2>
      <nltcs:v3>0</nltcs:v3>
   </nltcs:QNs>
</MolecularState>
```

XSAMS – Example: a molecular state of H2O

```
<RadiativeTransition methodRef="MEXP" sourceRef="B_HITRAN2008">
   <EnergyWavelength>
      <Wavenumber>
         <Experimental sourceRef="B_HIT-H2O-nu-30">
            <Value units="1/cm">1000.288940</Value>
            <Accuracy>1.000e-02</Accuracy>
         </Experimental>
      </Wavenumber>
   </EnergyWavelength>
   <InitialStateRef>S145-H2O-1</InitialStateRef>
   <FinalStateRef>S148-H2O-1</FinalStateRef>
   <Probability>
      <TransitionProbabilityA sourceRef="B_HIT-H2O-S-18">
         <Value units="1/s">2.335e-02</Value>
         <Accuracy>2.3e-03</Accuracy>
      </TransitionProbabilityA>
      <Multipole>E1</Multipole>
   </Probability>
   ...
</RadiativeTransition>
```

Advantages of Relational DB / XSAMS

- Easily extensible, for example:
  - more complex molecular states,
  - parameters for multiple lineshapes (Voigt, Galatry, …),
  - line-mixing effects;
- Data provenance:
  - each item of data can be given a source,
  - each data set requested from the online database can be given a timestamp and reproduced at a later time;
- Easy to **validate** the data …

Disadvantages of XSAMS:

- Extremely *verbose*: typically 50× larger file sizes;
- More computational power required to write and parse XML than “fixed” formats;
- Doesn’t play nicely with Fortran (yet).

But:

- Compresses well typically 50×!
- Can be transformed into other formats.

HITRAN data validation

- Introduction of a data model gives meaning to each item of data:
  - Can validate the quantum numbers assigned to each state (*e.g.* ensure $J \geq K$),
  - Can verify transitions obey certain selection rules (*e.g.* on parity: $+ \leftrightarrow -$ for electric dipole transitions);

- States are stored separately from Transitions:
  - Can verify that the same state is always given the same energy.

HITRAN data validation example: H_2S

HITRAN .par format:

```
311 2504.134000 1.884E-25 2.247E-02.07400.158 1699.03090.75-.
003000           0 2 0           0 0 0 13 12  1        13 31  2
452203 5 5 6 4 0 2    81.0    81.0
```

Lower state in XSAMS format:

```
<nltcs:QNs>
   <nltcs:ElecStateLabel>X</nltcs:ElecStateLabel>
   <nltcs:J>13</nltcs:J>
   <nltcs:Ka>31</nltcs:Ka>
   <nltcs:Kc>2</nltcs:Kc>
   <nltcs:v1>0</nltcs:v1>
   <nltcs:v2>0</nltcs:v2>
   <nltcs:v3>0</nltcs:v3>
</nltcs:QNs>
```

HITRAN data validation example: NH_3

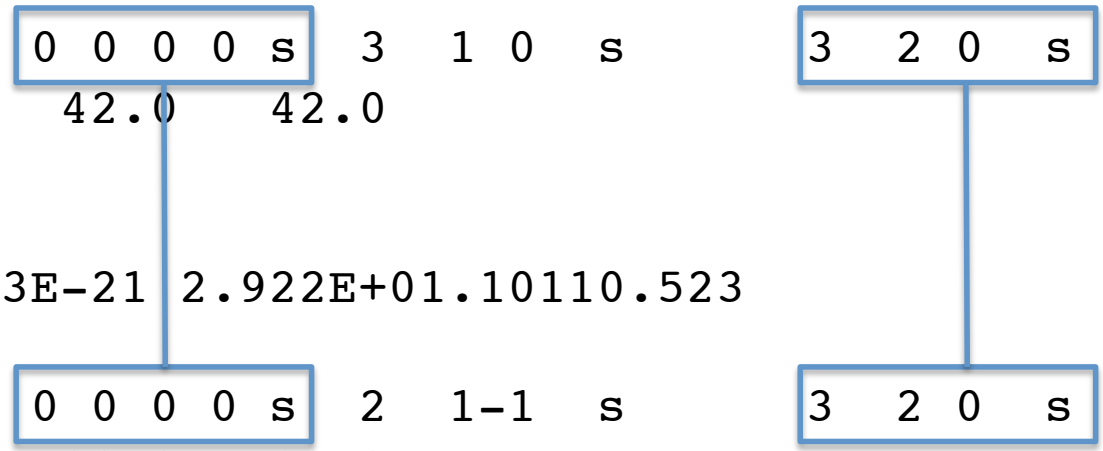
Two transitions in HITRAN .par format:

```
111   11.114541 1.549E-27 1.878E-10.10110.366
104.42210.700.000000
      0 0 0 0 s      0 0 0 0 s  3  1 0  s      3  2 0  s
542000 8 8 2 0 0 0    42.0   42.0


111 5007.100100 7.703E-21 2.922E+01.10110.523
256.38340.700.000000
      0 0 1 1 s      0 0 0 0 s  2  1-1  s      3  2 0  s
452000 4 3 2 0 0 0    30.0   42.0
```

HITRAN data validation example: NH_3

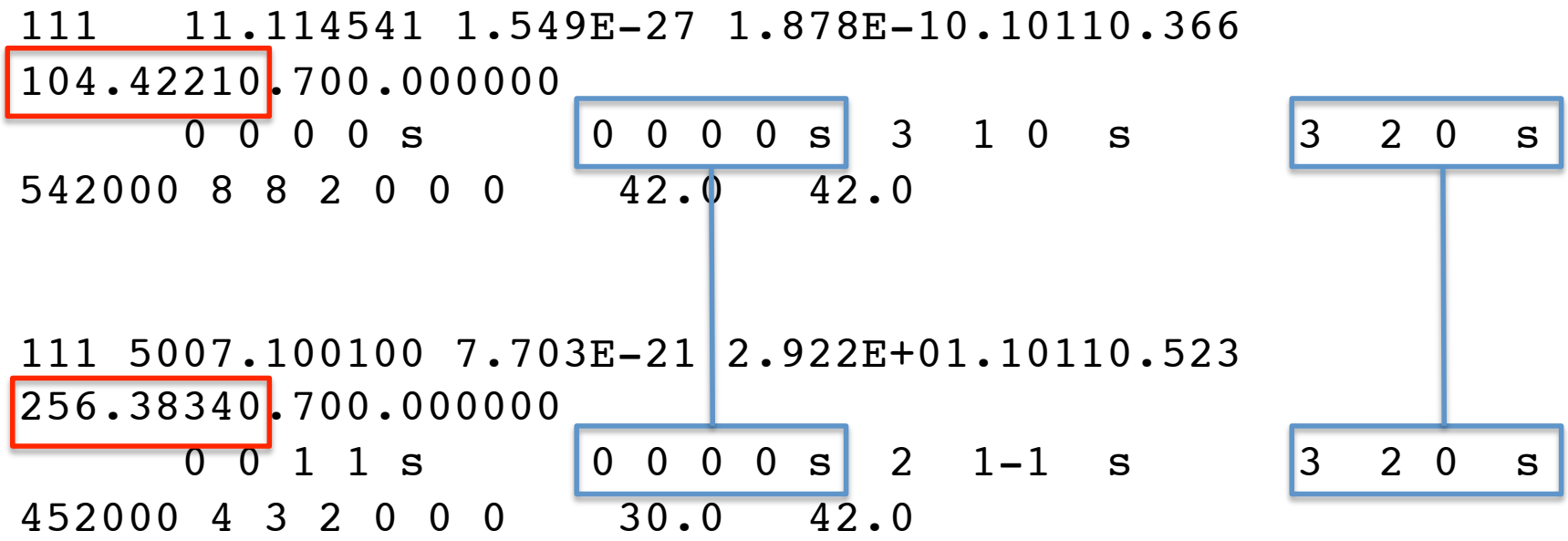
Two transitions in HITRAN .par format:

```
111   11.114541 1.549E-27 1.878E-10.10110.366
104.42210.700.000000
      0 0 0 0 s      0 0 0 0 s  3  1 0  s      3  2 0  s
542000 8 8 2 0 0 0     42.0   42.0

111 5007.100100 7.703E-21 2.922E+01.10110.523
256.38340.700.000000
      0 0 1 1 s      0 0 0 0 s  2  1-1  s      3  2 0  s
452000 4 3 2 0 0 0     30.0   42.0
```



HITRAN data validation example: NH_3

Two transitions in HITRAN .par format:



Inconsistencies identified

- Many examples of the same state being given different energies (affects the transition intensity temperature-dependence)
- NH_3
  - 6 states have $K > J$
  - 933 lines have inconsistent inversion symmetry labels (*a* and *s*)
- OH
  - 1096 states show an incorrect correlation of Hund's case (a) and case (b) quantum numbers (also for NO)
- H_2S
  - 53 states with $K_a > J$
- HOCl
  - 2104 states do not have $K_a + K_c = J$ or $J+1$

Acknowledgements

- Prof Jonathan Tennyson, UCL
- Dr Peter Yuen, MSSL
- VAMDC consortium:

VAMDC is funded under the "Combination of Collaborative Projects and Coordination and Support Actions" Funding Scheme of The Seventh Framework Program. Call topic: INFRA-2008-1.2.2 Scientific Data Infrastructure. Grant Agreement number: 239108