XML format for an NMRD data base?

Preliminary considerations and call for suggestions

Ing.Dr.Stanislav Sykora, August 5, 2002

Introduction

This Notes faces two distinct issues:

1. A proposal for the establishment of an NMRD **data base**.
2. A proposal for the adoption of the **XML** (**E**xtended **M**arkup **L**anguage) for encoding the data base entries.

A well structured data base format is a prerequisite for:

- Legible, standardized **encoding NMRD data** which would also guarantee their integrity.
- **Sharing data** among various research groups
- **Sharing acquisition procedures and parameters** across different instrumental platforms (commercial, home-made, various models of the same manufacturer, etc).

The data base should also enable its users to

- Convert individual entries into common format documents (doc, .pdf, .xls, .html, etc).
- Apply SQL queries to the whole data base.
- Edit the data base
- Cross-editing and merging of different data bases
- Importing individual entries into the FFC-NMR software (conversion to .sdf format).
- Automatic export of data from the FFC-NMR software into a data base.

An ad-hoc software doing all these things would be very costly and tedious to write and maintain. However, the problem has already been addressed on a much more general level and found a satisfactory (though still provisional) response in the XML language aimed specifically at handling documents containing complex data structures and references to external data bases.

It is therefore essential to make the data-base proposal compatible with the requirements of the XML standard[1] which is based on three types of scripts:

1. A data-definition (.dtd) file listing the data items of an entry and their characteristics.
2. The 'translator' files (.xsl) which describe how to convert an entry into a standard document of a particular type. One needs as many translators as their are target document formats.
3. Any number of actual entries (.xml files)

Once the logical contents of an data base entry have been defined, items (1) and (2) should be prepared by a central Authority (e.g., the instrument manufacturer) following the well-defined rules laid down by the W3C commission which develops and maintains the XML standard (the same one which maintains HTML).

Thereafter, individual items (3) can be written directly by the data base Users. The scripting becomes simple since the Users are free from any formatting considerations and need to concentrate exclusively at the actual contents (a software utility could further simplify the process).

The full data base would then consist of little more than a disk directory with any number .xml files. It would be automatically searchable by any search engine using the standard SQL language.

This Note is concerned with the very first step towards an NMRD data base, i.e., the definition of the **required and optional logical data items of an NMRD profile 'descriptor'** which might constitute a single NMRD data base entry.

The Author hereby calls the whole community for help. If you have any clear ideas what more (or less) should the would-be standard contain, please, let me know (sykora@ebyte.it).

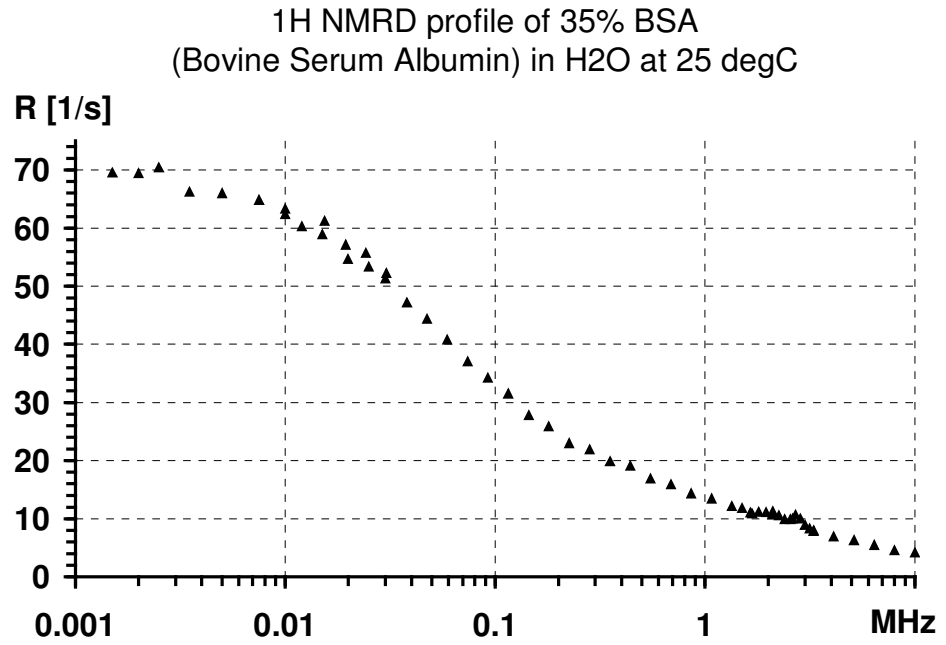
[1] Still more powerful concepts are presently in evolution. The proposed XMLS standard, for example, can incorporate data structures of amazing complexity.However, automatic conversion from older standards to newer ones is likely to eliminate any problem of early obsolescence.

Characterization of an *NMRD profile*

An NMRD profile measured by an FFC-NMR instrument, exemplified by the Figure and Table below[2], is essentially a scatter graph $G[B_{rlx},R_1]$ of R_1 against B_{rlx}, where:

- R_1 is the longitudinal relaxation rate, equal to the inverse of T_1, the longitudinal relaxation time.
- B_{rlx} is the magnetic field at which R_1 has been measured.

Example:



B_{rlx}	R_1	R_1 err	B_{rlx}	R_1	R_1 err	B_{rlx}	R_1
10	4.219	0.051	8	4.604	0.053	6.401	5.504
5.121	6.295	0.060	4.0967	6.926	0.063	3.2994	7.986
3.2782	7.88	etc	3.1506	8.322	etc	2.9998	8.9
2.8509	10.043		2.7005	10.681		2.6225	10.006
2.5508	9.901		2.4005	9.942		2.2499	10.561
2.0996	11.313		2.098	10.905		2.098	10.732
1.95	11.165		1.8003	11.187		1.5004	11.866
1.6784	10.946		1.6497	11.081		1.3422	12.217
1.0742	13.479		0.8594	14.348		0.6878	15.919
0.5502	16.911		0.4405	19.117		0.3524	19.901
0.2819	21.942		0.2253	23.022		0.1804	25.891
0.1445	27.825		0.1155	31.537		0.0922	34.256
0.0739	37.069		0.0591	40.776		0.0474	44.39
0.038	47.212		0.0303	52.266		0.0242	55.747
0.0194	57.155		0.0154	61.275		0.01	63.352
0.0199	54.701		0.01	62.426		0.005	66.02
0.015	58.91		0.002	69.466		0.0035	66.252
0.0075	64.865		0.012	60.285		0.025	53.357
0.03	51.354		0.0025	70.433		0.0015	69.544

[2] Sample origin: Bertil Halle, University of Lund, and Venu Kandadai, University of Hyderabad. The sample has been measured at Stelar as part of a preliminary instrument evaluation.

Introductory Notes on the example

Use of T_1 instead of R_1

Some Authors prefer to use T_1 instead of R_1 (probably because of historic reasons). However, R_1 is expected to be nearly additive with respect to various contributing relaxation mechanisms while, evidently, T_1 is not. Consequently, R_1 should be preferred and the use of T_1 should be discouraged.

Brlx units

At present, the magnetic field values are usually specified using the Larmor frequency of ^{1}H nuclides rather than the SI unit (Tesla). The conversion factor (there is a strictly linear) is 1 T ~ 42.57744 MHz. However, it would be better to present the graphs using a dual horizontal scale - one in SI units and one in Larmor frequency of *the measured nucleus*.

Brlx scale

The profile graphs are generally plotted using a logarithmic axis for B_{rlx} and a linear axis for R_1. Ideally, a profile should cover several decades of B_{rlx} values. The horizontal axis of the graphs can be to some extent standardized since at present it is hardly possible to measure reliable R1 values below 1 kHz since - from 10kHz down - the interference of environmental magnetic fields becomes rapidly a limiting factor. On the other side, 1 GHz is the absolute maximum for all currently available NMR-grade magnets. Consequently, when Brlx is expressed in ^{1}H Larmor frequency, $\log(B_{rlx})$ always lies in the interval [3,9]. For the commercially available Stelar instruments, the range of [3,7] is sufficient.

Error elipses

Ideally, each data point in the profile should be accompanied by a error estimates for both co-ordinates (the probable experimental error in R_1 and B_{rlx}). The two errors are *certainly uncorrelated* so that the resulting *error-ellipsis* has its axes aligned with the principal axes of the graph and can be represented by two orthogonal *error bars*. The Stelar software provides an estimate of the experimental error of R_1. The uncertainty of B_{rlx} is an instrumental characteristic for which there are at present no available data. The field noise of the Stelar instruments is of the order of 500 Hz (rms) - negligible for B_{rlx} values above 100 kHz but possibly influential at lower values. Procedures for an experimental determination of the B_{rlx} error bar are presently under study.

Multiple profiles

In some cases it is either desirable or even indispensable to resort to more than one curve within the same graph. A common situation of this type includes measurement of the same sample at several temperatures. For the purposes of a data base, one could in this case consider the distinct curves as separate (though related) entries. However, when the relaxation curves are measured for a large number of temperatures, the resulting graph $G[B_{rlx},Temp,R_1]$ describes a 2D *relaxation surface* which is of considerable value and should be viewed as a single object. Another case regards multi-exponential decays with multiple R_1 values. In this case would be highly inappropriate to spread the individual curves among multiple entries.

Additional data

In order to be of any use, an NMRD profile should be accompanied by additional data, some of which are indispensable (doubly underlined), some highly desirable (underlined) and some optional (italics). For archival purposes it is necessary to *keep the data down to a reasonable minimum* (the total number of parameters in the FFC-NMR program is over 200) *which describes the measurement in sufficient detail to make it reproducible on all instruments*.

Categories of data required in an NMRD data base

1) Sample description

- Assigned name
- *Description or chemical formula* (when applicable)
- *Origin*
- *Solvent* (when applicable) and c*oncentration* (when applicable)
- *Pre-treatment* (when important)
- Size (diameter and height in mm, e.g., d10,h15)
- Note

2) Thermodynamic and chemical state variables

- Sample temperature (preferably in degrees Kelvin) is of extreme importance. The dependence of R_1 on sample temperature is almost always quite strong so that statements like 'room temperature' are not acceptable.
- *Sample pressure* (in kPa) is considered much less important than sample temperature since the pressure dependence of R_1 is expected to be very weak. Whether this is really a universal rule is actually not known. At present there are no variable-pressure FFC probes and profiles are measured at 'normal' pressure which seems to be a good enough spec.
- Chemical state parameters:
  - pH (if applicable)
  - *Ionic force* (if applicable)

3) Data acquisition parameters

- Measured nuclide (e.g., ^{1}H, ^{13}C, ^{23}Na, …). Due to its high sensitivity, the nucleus presently most often measured is proton (^{1}H) but all Stelar instruments are multi-nuclear and NMRD profiles for nuclei other than ^{1}H can be measured without particular difficulties.
- Measurement technique described by means of a multi-parameter structure the form of which shall eventually become rigorously codified in an ASCII string. As a minimum, it shall contain the following parameters:
  - Low-field sequence (e.g., Balanced Pre-Polarized, Balanced Inversion Recovery, …)
  - High field sequence (e.g., Balanced Non-Polarized, Balanced Inversion Recovery, …)
  - Sequence-crossover field value (examples: 4 MHz, 0 MHz)
  - Signal detection method (FID, Echo, CPMG, other)
  - Note for optional data (e.g., sampling rate and low-frequency filter setting)
- Number and distribution of tau values. Each point of the NMRD profile is the result of an exponential (or multi-exponential) fit of a set of elementary measurements (blocks) in which there is a variable delay interval called “tau”. The number of the tau values and their distribution type may be of importance and should be specified (e.g., 16, Linear)
- Polarization field (e.g., 12 MHz)
- Polarization time (e.g., 0.2 sec)
- Acquisition field (e.g., 9.25 MHz)
- Magnet switching times (e.g., 2 ms / Optimized)
- Arrayed parameter (e.g., TEMP)
- Note

4) Data evaluation parameters

- Used data (real parts, magnitudes, signed magnitudes)
- Data reduction method (window average, FID analysis of some kind, CPMG fit)
- Relaxation decay fitting method (mono-, bi-, tri- exponential fit, other)
- Note

5) Instrument and operator characteristics

- Instrument description (examples: 'Stelar FFC-Spinmaster II', 'home made'). NMRD profiles should be independent of the instrument on which they were measured. One of the important purposes of a data base is establishing the degree to which this really true. It is for this reason that a description of the instrument and its basic characteristics is desirable.
- Operator
- Notes

6) Data base entry parameters

- In a data base, especially one managed on-line, one should be able to retrace the data to their original source, i.e., their Author. This is essential for citation purposes, Author rights protection and data validation (protection against hoaxes). The following fields should be therefore included:
  - Date
  - Author name and affiliation
  - Contact (address, e-mail or other)
  - Notes

Example of a hypothetical NMRD profile data-base entry

Note: each individual **section** of the data-base entry is specified by its name in square brackets, followed by the individual items (the required ones are in red color). The example describes two profiles of the same sample measured at two different temperatures (arrayed parameter = TEMP). The numeric values of the data are partly invented and do apply to the profile shown above.

[Sample description section]	
Assigned name	BSA #13
Description	Bovine Serum Albumin
Date of measurement	10/03/99
Origin	University of Lund
Solvent	H2O
Concentration	35%
Pre-treatement	Dialyzed, non-denatured
Size	id 9mm, h 12mm
Note	Very pure sample prepared by Dr.Finechemist
[Thermodynamic and chemical state variables]	
Sample temperature	arrayed
Sample pressure	normal
pH	6.3
Ionic force	n.a.
[Data acquisition parameters]	
Measured nuclide	1H
Measurement technique	PP,NP,4,FID /Macro Profile/
Tau values	16, 3*T1MX, 1 ms
Polarization field	12 MHz
Polarization time	0.2 s
Acquisition field	9.25 MHz
Switching time	3 ms
Recycle delay	0.2 s
Number of scans	4
Excitation parameters	PW90= 7 us, RINH=10 us
Basic acquisition parameters	SW=200 kHz, FLTR=90 kHz, BS=32, ACQD=15us
Arrayed parameter	TEMP
Notes	MTFC=650
[Data evaluation parameters]	
Used data	Magnitudes
Data reduction method	Window average, points 6 - end
Decay fitting method	Mono-exponential fit
Note	All block used
[Instrument characteristics]	
Instrument	Stelar FFC Spinmaster I
Operator	Stan Sykora
Note	0.25T, 2-layer magnet
[Data base entry parameters]	
Date	25/05/1999
Author name	Venu Kandadai
Affiliation	University of Hyderabad
Contact	venu@everest.edu
Notes	

[Data: Brlx, R1, fitting error]

```
  TEMP = 301.5 K
  START
  10, 4.219, 1.23
  5.121, 6.295, 0.70
  3.2782, 7.88, 1.75
  2.8509, 10.043, 0.33
  2.5508, 9.901, 0.98
  2.0996, 11.313, 1.02
  etc
  0.010, 63.352, 0.86
  END

TEMP = 311.5 K
  START
  10, 4.219, 1.23
  5.121, 6.295, 0.70
  3.2782, 7.88, 1.75
  2.8509, 10.043, 0.33
  2.5508, 9.901, 0.98
  2.0996, 11.313, 1.02
  etc
  0.010, 63.352, 0.86
  END

END
```