LA-UR- 09-06149

Approved for public release; distribution is unlimited.

Title:	Long-Term Data Archiving
Author(s):	D.S. Moore
Intended for:	Analytical and Bioanalytical Chemistry

Form 836 (7/06)

Long-Term Data Archiving

David S. Moore

Shock and Detonation Physics, DE-9, MS-P952, Los Alamos National Laboratory

Los Alamos, NM 87545 USA

Tel.: +1-505-665-6089, Fax: + 1-505-667-6372

moored@lanl.gov

Long term data archiving has much value for chemists, not only to retain access to research and product development records, but also to enable new developments and new discoveries. There are some recent regulatory requirements (e.g., FDA 21 CFR Part 11), but good science and good business both benefit regardless. A particular example of the benefits of and need for long term data archiving is the management of data from spectroscopic laboratory instruments. The sheer amount of spectroscopic data is increasing at a scary rate, and the pressures to archive come from the expense to create the data (or recreate it if it is lost) as well as its high information content. The goal of long-term data archiving is to save and organize instrument data files as well as any needed meta data (such as sample ID, LIMS information, operator, date, time, instrument conditions, sample type, excitation details, environmental parameters, etc.). This editorial explores the issues involved in long-term data archiving using the example of Raman spectral databases. There are at present several such databases, including common data format libraries and proprietary libraries. However, such databases and libraries should ultimately satisfy stringent criteria for long term data archiving, including readability for long times into the future, robustness to changes in computer hardware and operating systems, and use of public domain data formats. The latter criterion implies the data format should be platform independent and the tools to create the data format should be easily and publicly obtainable or developable. Several examples of attempts at spectral libraries exist, such as the ASTM ANDI format, and the JCAMP-DX format. On the other hand, proprietary library spectra can be exchanged and manipulated using proprietary tools. As the above examples have deficiencies according to the three long term data archiving criteria, Extensible Markup Language (XML; a product of the World Wide Web Consortium, an independent standards body) as a new data interchange tool is being investigated and implemented.

In order to facilitate data archiving, Raman data needs calibration as well as some other kinds of data treatment. Figure 1 illustrates schematically the present situation for Raman data calibration in the world-wide Raman spectroscopy community, and presents some of the terminology used below.

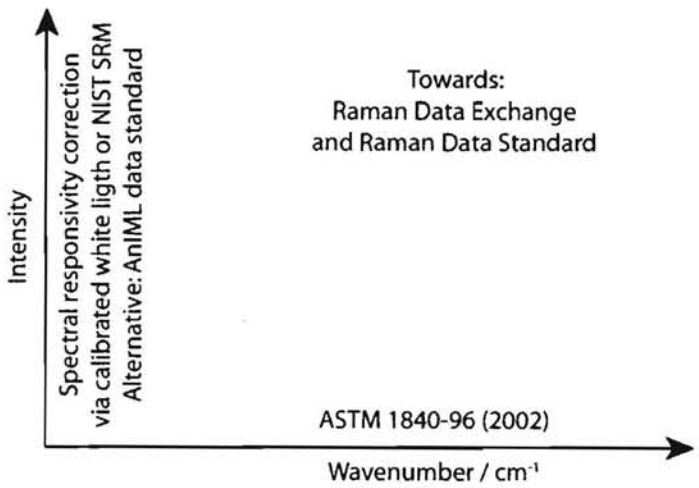

Figure 1. Present status of the Raman data standard as well as abscissa and ordinate calibrations.

Figure 2 shows three Raman spectra of polystyrene obtained using three different compact Raman systems. These spectra and the concepts in Fig. 1 will be used in the discussion below to illustrate the various facets of Raman data calibration.

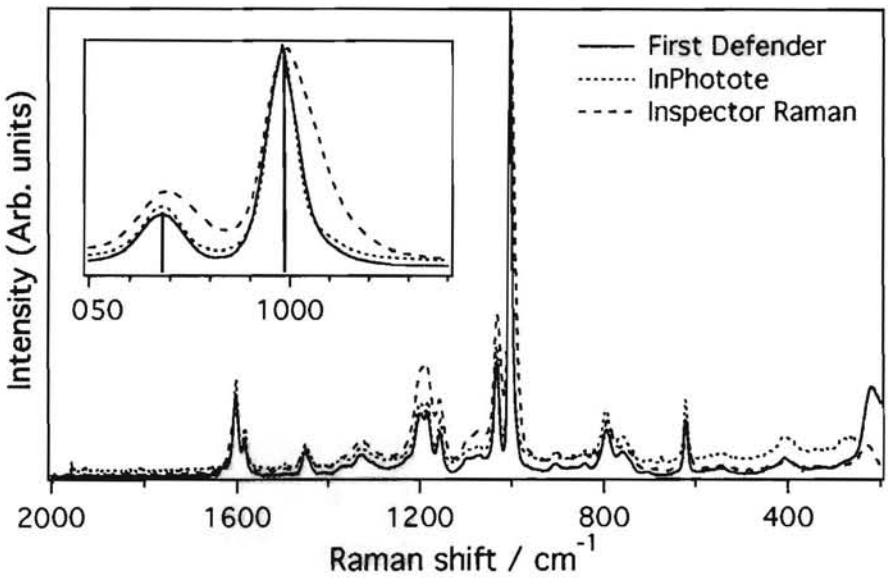

Figure 2: Raman spectra of polystyrene obtained using (solid line) First Defender (Ahura Scientific), (dotted line) InPhotote (Inphotonics, Inc.), and (dashed line) Inspector Raman (DeltaNu, LLC). The solid vertical lines in the inset are located at the ASTM-E-1840 calibrated wavenumbers and relative intensities.

In order to enable the most accurate comparison of Raman spectra obtained from various instruments, abscissa (wavenumber axis) calibration must be performed, and some method of accounting for the system spectral responsivity should be used to correct the ordinate. For

abscissa calibration, the ASTM E-1840 guide recommends several means [1]. The ASTM utilized a round robin methodology to determine the peak wavenumbers for eight different readily available materials having no known polymorphs, selected to cover a wide wavenumber range (85 to 3327 cm^{-1}) for both solids and liquids. The standard error for the peaks is stated to be < 1.0 cm^{-1}. Often, Raman instruments have some kind of calibration routines built in, which involve measurement of the excitation laser wavelength and calibration of the detector pixel versus wavelength using an atomic line source. The results of this type of calibration should be verified using at least one of the approved ASTM-E-1840 guide materials before archiving abscissa wavenumbers for any Raman spectral features.

The spectra in Figure 2 were abscissa calibrated according to each instrument vendor's instructions or routines. The inset shows the recommended polystyrene wavenumber values as vertical lines. Two of the instruments produced spectra with calibrations better than 0.4 cm^{-1} compared to ASTM-E-1840, and the other instrument's spectra were calibrated to ~ 1 cm^{-1}, the standard error for the ASTM recommendations.

One abscissa parameter not considered by the ASTM-E-1840, which has less to do with wavenumber calibration than with the ultimate ability of library search algorithms to identify materials, is the instrumental spectral resolution. The spectra in Figure 2 show that two of the instruments have very similar resolution (~5 cm^{-1}) but one has significantly poorer resolution (~8 cm^{-1}). The ASTM-E-1840 recommends peak wavenumbers, but some search routines are also influenced by line widths, so that library spectra obtained with instruments of different spectral resolution could result in identification errors. This issue still needs to be resolved, with one possible solution being deconvolution of the instrument spectral resolution function before archiving.

Correction of ordinate data is more problematic. The ν^4 factor (dependence of Raman cross section on excitation frequency) causes changes in Raman signal strength with excitation wavelength [2-4]. In addition, electronic pre-resonance and resonance effects can cause changes in the relative intensities of various Raman features depending on their associated chromophore. For a given selected excitation wavelength, however, the ordinate correction process is straightforward. The detected Raman intensities should be multiplied by a relative spectral responsivity correction curve obtained for the entire optical train from sample to detector. Two methods are available to measure the spectral responsivity curve. One is using a calibrated white light irradiance source. More recently, NIST has developed three certified luminous optical glasses (for different excitation wavelengths, 785, 532, and 488/514.5 nm) that emit a broad luminescence spectrum when illuminated with the Raman excitation laser [5]. To perform the correction, either the white light irradiance or the glass luminescence is recorded using the Raman system. Then the spectral irradiance value curve or the relative spectral luminescence curve (I_{SRM}) is used to obtain the spectral intensity correction curve C_{SRM} by dividing I_{SRM} by the spectrum measured by the Raman instrument S_{SRM}. The relative intensity corrected Raman spectrum is then obtained by multiplying the measured Raman spectrum by C_{SRM}.

Note the ordinate differences shown in Figure 2 for the polystyrene Raman spectrum obtained using three different portable Raman systems, which is indicative of the kinds of ordinate calibration problems that can occur. At low wavenumbers, one of the spectra has too high a baseline compared to the other two. In the middle wavenumber region, another of the instruments produces peaks that are too large compared to the other two, relative to the largest peak at 1001.4 cm^{-1}. The inset shows the relative intensities of two features near 1000 cm^{-1}, for which spectra from two instruments match very well, but the spectrum from the third instrument (the same one that has lower spectral resolution) is quite different. These differences apparently occur because of inadequate (or lack of) ordinate spectral responsivity correction, and could affect library search outcomes (see below).

There are a number of Raman spectral databases available [6]. For example, Raman spectra for a large number of minerals are available at the Caltech/University of Arizona RRUFF™ Project website (allows download of spectra in ASCII text pairs) [7]. Raman spectra of natural and synthetic pigments are available at the University College London web site [8]. The University of Siena has a Raman database (graphical spectra with listing of selected peaks) for minerals and inorganic materials [9]. A searchable database of molecular spectra (from NMR to IR and Raman) of organic compounds is available from the AIST in Japan [10]. Some commercial enterprises offer on-line Raman database searching, such as the FTIR/Raman search.com from Thermo Fisher Scientific. It uses the GRAMS/Al methodology and Thermo Scientific spectral databases [11]. In addition, a number of instrument manufacturers offer extensive proprietary Raman spectral databases.

Nevertheless, these databases do not satisfy the stringent criteria for long term data archiving, which include: readability for long times into the future, robustness to changes in computer hardware and operating systems, and use of public domain data formats. The latter criterion implies that data formats should be platform independent and the tools to create the data forms should be easily and publicly obtainable or developable. In the US, there is regulatory pressure to archive "accurate and complete" copies of electronic records provided to government agencies (FDA 21 CRF 11: Data Formats).

To date, none of the available Raman spectral databases abide by all long term data archiving criteria, yet there has been some progress. Some open source spectral library formats have been developed. The ASTM E01.25 Subcommittee on Laboratory Analytical Data Interchange Protocols and Information Management develops analytical data interchange protocols (ANDI protocols) to increase laboratory efficiency and productivity by facilitating the integration and use of data from multiple vendors' instruments. The IUPAC developed the JCAMP-DX data format for spectroscopic data [12]. JCAMP is completely ASCII based for simple transport and readability, and utilizes a fixed dictionary of tags, but suffers from numerical accuracy issues (translation of binary to decimal). Galactic developed the SPC data format primarily for optical spectroscopy, and has made its structure available in the public domain [13]. The ANDI and SPC

protocols suffer from not being human readable. None of the three are easily validated for formatting and content, nor extensible for future changes in equipment and analysis methods.

XML is not a file format, but rather a universal markup languange for exchanging structured documents and data on the Web with its own vocabulary and syntax, constrained by the recently adopted XML Schema Definition (XSD) language. It can be used to represent any data structure. The XML model provides the ability to link to other documents, called XML Schema, that describe exactly what each piece of data means and how it is related via XML language descriptions of a particular type of information and its storage. With the appropriate Schema, a software application can parse data from any XML data structure as well as determine that it is well formed. Therefore, XML documents can be externally validated for both content and syntax. XML data is human readable ASCII text, and the XML standard is public domain managed by the World Wide Web Consortium (W3C).

The latest implemented version of this idea is embodied in the Analytical Information Markup Language, ANiML, http://animl.sourceforge.net/ [14]. The project is a collaborative effort between many groups and individuals and is sanctioned by the ASTM under subcommittee E13.15. A variety of techniques have Definition Documents, which apply tight constraints to the flexible core and are in turn defined by the Technique Schema. The Open Source development is hosted on SourceForge. As of this writing, there is a Technique Definition Document for IR and UV/visible spectroscopic data, as well as sample parameters, but not yet for Raman spectroscopic data. A Web Blog is available to aid tracking of the development of software tools to take advantage of the new data standard – http://www.animltools.com/home [15].

References

1. ASTM International, **2002**, ASTM E1840-96(2002) Standard guide for Raman shift standards for spectrometer calibration
2. B. Schrader, D.S. Moore, **1997**, Pure Appl. Chem. 69:1451-1468
3. D.A. Long, **2002**, The Raman Effect: a unified treatment of the theory of Raman scattering by molecules, (Wiley, Chichester).
4. B. Schrader, **1995**, Infrared and Raman Spectroscopy, (VCH, Weinheim).
5. S.J. Choquette, E.S. Etz, W.S. Hurst, D.H. Blackburn, S.D. Leigh, **2007**, Appl. Spectrosc. 61:117-129
6. M.B. Denton, R.P. Sperline, J.H. Giles, D.A. Gilmore, C.J.S. Pommier, R.T. Downs, **2003**, Aust. J. Chem. 56:117-131
7. http://rruff.info/
8. http://www.chem.ucl.ac.uk/resources/raman/index.html

9. http://www.dst.unisi.it/geofluids/raman/spectrum_frame.htm
10. http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi?lang=eng
11. https://ftirsearch.com/default2.htm
12. http://www.jcamp-dx.org/
13. http://www.thermo.com/com/cda/resources/resources_detail/1,,112125,00.html
14. AnIML: Analytical Information Markup Language, **2008**, http:// animl.sourceforge.net/
15. Scimatic Software, **2008**, http://www.animltools.com/home