INTERNATIONAL STANDARD

ISO 22029

First edition
2003-12-15

Standard file format for spectral data exchange

Format standard de fichier pour échange de données spectrométriques

Reference number
ISO 22029:2003(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword iv
Introduction v
1 Scope 1
2 Format description 1
2.1 General 1
2.2 Required keywords 2
2.3 Spectral data 3
2.4 Optional keywords 4
2.5 Ending a file 6
3 Example of a data file in the EMSA spectral format 7
Bibliography 9

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 22029 was prepared by Technical Committee ISO/TC 202, *Microbeam analysis*.

Introduction

The original EMSA/MAS Standard File Format for Spectral Data Exchange was published in October 1991. Since then, advances in both microbeam analysis techniques and in PC technology have meant that this original standard is not fully able to meet modern requirements. The members of ISO/TC 202 (the International Organization for Standardization, Technical Committee for Microbeam Analysis) propose this updated version. Every effort has been made to only make those changes that improve or update the applicability of the standard, whilst minimizing incompatibility with the original version. The remit of TC 202 does not include surface analysis techniques, which are addressed by TC 201, and so references to these techniques have been removed from the original standard where necessary. The original document also included examples of coding and telecommunications protocols. Since these are now largely outdated, and not integral to the formatting of the data, these have also been removed.

This International Standard is presented as the text of the original, with modifications shown in *italic text*. It is further noted that one of the originating societies (EMSA) has modified its name since the original document was published. The society is now officially known as the “Microscopy Society of America” or MSA, the term “Electron” having been dropped to more fully indicate the work and interest of the membership of the society in all forms of microscopy.

The virtues of a single standard data format have been admirably related by various authors [1], [2], [3], [4]. It would often be convenient, after visiting another laboratory in order to use a different type of microanalytical spectrometer, to be able to return to one's own laboratory to analyse the data, or for a laboratory to be able to send a spectrum to another group at another location for analysis on their computer. A common format would also enable test spectra to be transported between data acquisition systems, in order to compare different data-analysis routines, and would give users greater choice of analysis procedure, based on commercial or public-domain software.

Obviously, an ideal solution would be for the manufacturers to represent data in a standard format, but they are unlikely to agree on this without some direction from their customers (the microanalysis community). Therefore it is highly desirable for EMSA and MAS to proceed with adoption of a standard format. Such a format does not preclude any research group or manufacturer from having their own, possibly proprietary, format. Spectral data can be stored internally in any format, as long as there is an option to convert it to the external standard (and vice versa) for the purposes of exchange. We believe that a standard format should pose the following attributes.

1) It should be capable of representing the data exactly (without altering the scientific content).

2) The format should be simple and easy to use.

3) It must NOT be tied to any particular computer, programming language or operating system. It should work on a large number of computers of all sizes, although we cannot guarantee that it will work on all possible computers.

4) The format should be both human- and machine (computer)-readable.

5) It should be compatible with existing electronic communication networks such as (but not necessarily limited to) BITNET, ARPANET, and with the phone system (using modems). Future networks will likely retain compatibility with these.

6) The format should support spectra of interest to the EMSA/MAS community (such as XEDS, EELS, AES, etc.) and should be flexible enough to accommodate future data sets, not yet specified.

7) Each file should contain enough information to uniquely identify the type and origin of the spectral data and to reconstruct its significance.

8) Where possible, the format should be compatible with various commercial data plotting or analysis programs (i.e. spreadsheets or graphical analysis packages).

9) The proposed format need not be the most efficient storage mechanism. Its primary goals, stated above, will generally prevent storage efficiency. If anything, this format will err on the side of simplicity and ease of use.

The format originally employed by the Electron Microscopy and Microanalysis Public Domain Library (EMMPDL) at Argonne [1] has the virtue of simplicity, but is too rigid for general use. A recent revision [5] corrected some inadequacies, but a more serious re-examination is in order. The format proposed by a previous EMSA Task Force [3], [4] addresses many of the problems but is thought by some microscopists to be too complicated for everyday use. The VMAS format, whose description [6] runs to 60 pages, is also too complex for our perceived purpose. A format, named JCAMP-DX, used by the infrared spectroscopy community [2] is specific and detailed but is somewhat off-target for the spectroscopies of interest to our community.

The format proposed here follows JCAMP-DX in many ways, but is less complicated and has features tailored to X-ray, energy-loss and Auger spectroscopies. We circulated a preliminary version of this proposal to several manufacturers of XEDS systems and have received comments and suggestions, many of which have been incorporated into this document.

The companion problem of a standard format for digital image storage is similar to that of spectral data, but is sufficiently different to warrant its own standard. Whereas most spectra are sufficiently compact that they can be stored in a human-readable form, image data are usually so extensive as to require storage of “raw” binary numbers. There exist formats for image storage that are in widespread use. One of these (probably TIFF) should be endorsed by EMSA and MAS, allowing the microanalysis community to take advantage of the large amount of commercial and public-domain software already available.

Standard file format for spectral data exchange

1 Scope

This International Standard presents a simple format for the exchange of digital spectral data that has been designated as an EMSA/MAS standard. This format is readable by both humans and computers and is suitable for transmission through various electronic networks (BITNET, ARPANET), the phone system (with modems) or on physical computer storage devices (such as floppy disks). The format is not tied to any one computer, programming language or computer operating system.

The adoption of a standard format would enable different laboratories to freely exchange spectral data, and would help to standardize data analysis software. If equipment manufacturers were to support a common format, the microscopy and microanalysis community would avoid duplicated effort in writing data-analysis software.

2 Format description

2.1 General

The general structure of the data file format can be summarized as a simple sequential ASCII (text) file. It begins with a series of header lines which serve to define the characteristics of the spectrum. These header lines are identified by unique keyword fields which occupy the first 15 positions of each line, followed by a data field. These are described in detail below. After the header lines, a keyword indicates the start of data, the data then follow on successive lines in a manner which is defined explicitly within the header. Finally, after all the data are presented, an end of data keyword indicates the data set is complete. This is diagrammatically illustrated below.

Header lines
*
*
Successive lines beginning with EMSA/MAS defined keywords, some of which are required and some are optional
*
*
Start of data keyword
*
*
Experimental data
*
*
End of data keyword

In general, each line of the file either contains a keyword and its associated value or spectral data. All header lines are readily identified as they each begin with “#” in the first character field or column. This symbol demarks the start of a 13 character keyword field, providing descriptive information about the data followed by an associated value. EMSA/MAS-defined keywords (whose definition may be changed only by EMSA/MAS) begin with a single # and occupy the first 13 columns (characters) of each header line. The keyword itself consists of at most twelve characters, which directly follow the #, shorter keywords may be employed and any remaining spaces following the defined keyword may be filled with descriptive text such as unit designation for ease of legibility when the file is printed (see examples in Table 1). If a position in the keyword field is not used,

it must be filled with a space character. The keywords are not case-sensitive, so that #Xunits is interpreted the same as #XUNITS.

The 14th and 15th character positions (columns) in each header line are occupied by keyword/value field separators, which consist of a colon followed by a space. The value associated with each keyword starts in column 16 and may be either textual or numeric as defined by the keyword. Each line of the file may contain no more than 79 characters (for compatibility with the largest number of computers and computer networks, and for general legibility on standard width CRT screens). Since the keyword and its separator occupy the first 15 positions, this means that all remaining information following the keyword is limited to a maximum of 64 (= 79 – 15) character positions. The end of line indicator is a carriage return, linefeed combination (<CR><LF>).

The only characters allowed in the file are the space (ASCII character 32), carriage return (ASCII character 13) and linefeed (ASCII character 10) characters, plus the printable ASCII character set given below.

!"#$%&'()*+,-./0123456789:;<=>? @ (ASCII characters 33 - 64)
ABCDEFGHIJKLMNOPQRSTUVWXYZ (ASCII characters 65 - 90)
abcdefghijklmnopqrstuvwxyz (ASCII characters 97 - 122)
[\] `{|}~ (ASCII characters 91 - 96, ASCII 123-126)

Horizontal TAB characters are NOT permitted in this file format as a substitute for spaces or commas. Examples of keywords, separators and data can be found in Table 1, and are further detailed below. *The file should be named using the extension .TXT, .EMSA, or .MSA.*

2.2 Required keywords

The following keywords are required and must appear at the beginning of the file, in the order specified below. Although some of these may appear arbitrary, in our judgment they fulfil a long-term need. After several years of students and outside users passing through a laboratory, the result can be a vast number of data files of unknown origin. Unless some adequate form of labelling is imposed from the start, many worthwhile data files are lost and useless and data sit on a disk taking up valuable space. With the following minimum subset of keywords, it should be possible to reconstruct the significance of most spectra.

Note that there must be exactly one of each required keyword, except for the keyword #TITLE which must appear at least once but may, at the user's discretion, appear more than once to provide an extended length title.

Required Keywords:

#FORMAT = Character string identifies this format as "EMSA/MAS Spectral Data File"

#VERSION = File format version number (*TC 202 v*1.0 for this implementation)

#TITLE = Gives a short description of the spectra. One or more per file. Max. = 64 characters

#DATE = The calendar day-month-year in which the spectra was recorded, in the form: DD-MMM-YYYY.

#TIME = The time of day at which the spectrum was recorded, in 24-hour format: HH:MM.

#OWNER = The name of the person who recorded the spectrum.

#NPOINTS = Total number of data points in X & Y data arrays
1. $\leqslant$ NPOINTS. *The original EMSA standard set an upper limit of 4096. This was considered to be too restrictive and unnecessary for modern computer systems.*

#NCOLUMNS	=	Number of columns of data 1. ⩽ NCOLUMNS < 5. when DATATYPE = Y 1. ⩽ NCOLUMNS < 3. when DATATYPE = XY *It is strongly recommended that values > 1 are not used since multicolumn formats cause problems when importing data into spreadsheet packages. They were originally intended to make more efficient use of space on teletype printouts of data files.*
#XUNITS	=	Units for X-axis data, e.g.: eV
#YUNITS	=	Units for Y-axis data, e.g.: counts.
#DATATYPE	=	Method in which the data values are stored as: Y axis only values or X,Y data pairs. The current options are the characters Y and XY.
#XPERCHAN	=	The number of X-axis units per channel.
#OFFSET	=	A real (but possibly negative) number representing value of channel one in xunits.
#SPECTRUM	=	Indicates that the next line starts the spectral data.
#ENDOFDATA	=	Indicates the end of the data file.

2.3 Spectral data

The spectral data must be enclosed between the following keywords:

#SPECTRUM : Signifies the beginning of spectral data (on the next line).

#ENDOFDATA : Signifies the end of spectral data.

Between these keywords, the spectrum is listed in one of two ways, as specified by the #DATATYPE value.

In the case of spectra with equally-spaced x-values (equal increments per channel), up to #NCOLUMNS y-values may be given per line. For Y-axis data type 1 < #NCOLUMNS < 5. Each y-value is either a real number (containing a decimal point, even if there is no fractional component) or is expressed in exponential form (e.g. 3.142E+3), and is followed immediately by a comma. The corresponding x-values can be obtained from the specified values of #XPERCHAN and #OFFSET or #CHOFFSET. *Due to the problems posed by multi-column formats, for importing data into spreadsheet programs etc., it is strongly recommended that only a single column format (#NCOLUMNS = 1) be used.*

If the user prefers, data may be stored as (x,y) pairs of data points, at #NCOLUMNS per line. For XY-axis data type 1 < #NCOLUMNS < 3. The x- and y-values are expressed as real numbers, or in exponential format, and are separated by a comma. The (x,y) pairs themselves are separated by a comma followed by at least one blank space. *Due to the problems posed by multi-column formats, for importing data into spreadsheet programs etc., it is strongly recommended that only a single column format (#NCOLUMNS = 1) be used*

The reason for avoiding the use of integer numbers is that in some instances decimal values are generated, e.g., if energy-loss spectra have been scaled for normalization or to remove a gain change. If necessary, the decimal point and the comma can be removed with a text editor (using a global replace) to give integer values, whereas the reverse process may not be straightforward.

Numbers less than unity can be represented either with or without a zero preceding the decimal point (e.g. 0.1 or .1). In the case of negative numbers, there should be no spaces between the minus sign and the numerical value. We recommend that there should be no trailing spaces after a number (preceding the comma). The (x,y) option has been included to accommodate segmented spectra, containing gaps where y-data are not specified, and to allow for the possibility that the X-axis scale is non-linear. In addition, it makes the data compatible with most general-purpose graph-plotting software packages. In some future version of the format,