Designation: E 1947 – 98 (Reapproved 2004)

Standard Specification for Analytical Data Interchange Protocol for Chromatographic Data[1]

This standard is issued under the fixed designation E 1947; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This specification covers a standardized format for chromatographic data representation and a software vehicle to effect the transfer of chromatographic data between instrument data systems. This specification provides protocol designed to benefit users of analytical instruments and increase laboratory productivity and efficiency.

1.2 The protocol in this specification provides a standardized format for the creation of raw data files or results files. This standard format has the extension ".cdf" (derived from NetCDF). The contents of the file include typical header information like instrument, column, detector, and operator description followed by raw or processed data, or both. Once data have been written or converted to this protocol, they can be read and processed by software packages that support the protocol.

1.3 The software transfer vehicle used for the protocol in this specification is NetCDF, which was developed by the Unidata Program and is funded by the Division of Atmospheric Sciences of the National Science Foundation.[2]

1.4 The protocol in this specification is intended to (*1*) transfer data between various vendors' instrument systems, (*2*) provide LIMS communications, (*3*) link data to document processing applications, (*4*) link data to spreadsheet applications, and (*5*) archive analytical data, or a combination thereof. The protocol is a consistent, vendor independent data format that facilitates the analytical data interchange for these activities.

1.5 The protocol consists of:

1.5.1 This specification on chromatographic data, which gives the full definitions for each one of the generic chromatographic data elements used in implementation of the protocol. It defines the analytical information categories, which are a convenient way for sorting analytical data elements to make them easier to standardize.

1.5.2 Guide E 1948 on chromatographic data, which gives the full details on how to implement the content of the protocol using the public-domain NetCDF data interchange system. It includes a brief introduction to using NetCDF. It is intended for software implementors, not those wanting to understand the definitions of data in a chromatographic dataset.

1.5.3 NetCDF Users Guide.

2. Referenced Documents

2.1 *ASTM Standards:* [3]

E 1948 Guide for Analytical Data Interchange Protocol for Chromatographic Data

2.2 *Other Standard:*

NetCDF User's Guide[4]

2.3 *ISO Standards:*[5]

2014-1976 (E) Writing of Calendar Dates in All-Numeric Form

3307-1975 (E) Information Interchange—Representations of Time of the Day

4031-1978 (E) Information Interchange—Representations of Local Time Differentials

3. Terminology

3.1 *Definitions for Adminstrative Information Class*—These definitions are for those data elements that are implemented in the protocol. See Table 1.

3.1.1 *administrative-comments*—comments about the dataset identification of the experiment. This free test field is for anything in this information class that is not covered by the other data elements in this class.

3.1.2 *company-method-id*—internal method id of the sample analysis method used by the company.

---

[1] This specification is under the jurisdiction of ASTM Committee E01 on Analytical Chemistry for Metals, Ores and Related Materials and is the direct responsibility of Subcommittee E01.25 on Laboratory Data Interchange and Information Management.

Current edition approved April 1, 2004. Published May 2004. Originally approved in 1998. Last previous edition approved in 1998 as E 1947–98.

[2] For more information on the NetCDF standard, contact Unidata at www.unidata.ucar.edu.

[3] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

[4] Available from Russell K. Rew, Unidata Program Center, University Corporation for Atmospheric Research, P. O. Box 3000, Boulder, CO 80307-3000.

[5] Available from ISO, 1 Rue de Varembe, Case Postale 56, CH 1211, Geneve, Switzerland.

TABLE 1 Administrative Information Class

NOTE 1—Particular analytical information categories (C1, C2, C3, C4, or C5) are assigned to each data element under the Category column. The meaning of this category assignment is explained in Section 5.

NOTE 2—The Required column indicates whether a data element is required, and if required, for which categories. For example, M1234 indicates that that particular data element is required for any dataset that includes information from Category 1, 2, 3, or 4. M4 indicates that a data element is only required for Category 4 datasets.

NOTE 3—Unless otherwise specified, data elements are generally recorded to be their actual test values, instead of the nominal values that were used at the initiation of a test.

Data Element Name	Datatype	Category	Required
dataset-completeness	string	C1	M12345
protocol-template-revision	string	C1	M12345
netcdf-revision	string	C1	M12345
languages	string	C5	. . .
administrative-comments	string	C1 or C2	. . .
dataset-origin	string	C1	M5
dataset-owner	string	C1	. . .
dataset-date-time-stamp	string	C1	. . .
injection-date-time-stamp	string	C1	M12345
experiment-title	string	C1	. . .
operator-name	string	C1	M5
separation-experiment-type	spring	C1	. . .
company-method-name	string	C1	. . .
company-method-id	string	C1	. . .
pre-experiment-program-name	string	C5	. . .
post-experiment-program-name	string	C5	. . .
source-file-reference	string	C5	M5
error-log	string	C5	. . .

3.1.3 *company-method-name*—internal method name of the sample analysis method used by the company.

3.1.4 *dataset-completeness*—indicates which analytical information categories are contained in the dataset. The string should exactly list the category values, as appropriate, as one or more of the following "C1+C2+C3+C4+C5," in a string separated by plus (+) signs. This data element is used to check for completeness of the analytical dataset being transferred.

3.1.5 *dataset-date-time-stamp*—indicates the absolute time of dataset creation relative to Greenwich Mean Time. Expressed as the synthetic datetime given in the form: YYYYMMDDhhmmss±ffff.

3.1.5.1 *Discussion*—This is a synthesis of ISO 2014, ISO 3307, and ISO 4031, which compensates for local time variations.

3.1.5.2 *Discussion*—The time differential factor (ffff) expresses the hours and minutes between local time and the Coordinated Universal Time (UTC or Greenwich Mean Time, as disseminated by time signals), as defined in ISO 3307. The time differential factor (ffff) is represented by a four-digit number preceded by a plus (+) or a minus (-) sign, indicating the number of hour and minutes that local time differs from the UTC. Local times vary throughout the world from UTC by as much -1200 hours (west of the Greenwich Meridian) and by as much as +1300 hours (east of the Greenwich Meridian). When the time differential factor equals zero, this indicates a zero hour, zero minute, and zero second difference from Greenwich Mean Time.

3.1.5.3 *Discussion*—An example of a value for this date element would be: 1991,08,01,12:30:23-0500 or 19910801123023-0500. In human terms this is 12:30 PM on August 1, 1991 in New York City. Note that the -0500 hours is 5 full hours time behind Greenwich Mean Time. The ISO standards permit the use of separators as shown, if they are required to facilitate human understanding. However, separators are not required and consequently shall not be used to separate date and time for interchange among data processing systems.

3.1.5.4 *Discussion*—The numerical value for the month of the year is used, because this eliminates problems with the different month abbreviations used in different human languages.

3.1.6 *dataset-origin*—name of the organization, address, telephone number, electronic mail nodes, and names of individual contributors, including operator(s), and any other information as appropriate. This is where the dataset originated.

3.1.7 *dataset-owner*—name of the owner of a proprietary dataset. The person or organization named here is responsible for this field's accuracy. Copyrighted data should be indicated here.

3.1.8 *error-log*—information that serves as a log for failures of any type, such as instrument control, data acquisition, data processing or others.

3.1.9 *experiment-title*—user-readable, meaningful name for the experiment or test that is given by the scientist.

3.1.10 *injection-date-time-stamp*—indicates the absolute time of sample injection relative to Greenwich Mean Time. Expressed as the synthetic datetime given in the form: YYYYMMDDhhmmss⁺⁄₋ ffff. See *dataset-date-time-stamp* for details of the ISO standard definition of a date-time-stamp.

3.1.11 *languages*—optional list of natural (human) languages and programming languages delineated for processing by language tools.

3.1.11.1 *ISO-639-language*—indicated a language symbol and country code from Annex B and D of the ISO-639 Standard.

3.1.11.2 *other-language*—indicates the languages and dialect using a user-readable name; applies only for those languages and dialects not covered by ISO 639 (such as programming language).

3.1.12 *Netcdf-revision*—current revision level of the NetCDF data interchange system software being used for data transfer.

3.1.13 *operator-name*—name of the person who ran the experiment or test that generated the current dataset.

3.1.14 *post-test-program-name*—name of the program or subroutine that is run after the analytical test is finished.

3.1.15 *pre-test-program-name*—name of the program or subroutine that is run before the analytical test is finished.

3.1.16 *protocol-template-revision*—revision level of the template being used by implementors. This needs to be included to tell users which revision of E 1947 should be referenced for the exact definitions of terms and data elements used in a particular dataset.

3.1.17 *separation-experiment-type*—name of the separation experiment type. Select one of the types shown in the following list. The full name should be spelled out, rather than just

referencing the number. This requirement is to increase the readability of the datasets.

3.1.17.1 *Discussion*—Users are advised to be as specific as possible, although for simplicity, users should at least put "gas chromatography" for GC or "liquid chromatography" for LC to differentiate between these two most commonly used techniques.

Separation Experiment Types

Gas Chromatography
- Gas Liquid Chromatography
- Gas Solid Chromatography

Liquid Chromatography
- Normal Phase Liquid Chromatography
- Reversed Phase Liquid Chromatography
- Ion Exchange Liquid Chromatography
- Size Exclusion Liquid Chromatography
- Ion Pair Liquid Chromatography
- Other

Other Chromatography
- Supercritical Fluid Chromatography
- Thin Layer Chromatography
- Field Flow Fractionation
- Capillary Zone Electrophoresis

3.1.18 *source-file-reference*—adequate information to locate the original dataset. This information makes the dataset self-referenced for easier viewing and provides internal documentation for GLP-compliant systems.

3.1.18.1 *Discussion*—This data element should include the complete filename, including node name of the computer system. For UNIX this should include the full path name. For VAX/VMS this should include the node-name, device-name, directory-name, and file-name. The version number of the file (if applicable) should also be included. For personal computer networks this needs to be the server name and directory path.

3.1.18.2 *Discussion*—If the source file was a library file, this data element should contain the library name and serial number of the dataset.

3.2 *Definitions for Sample-Description Information Class*—This information class is comprised of nominal information about the sample. This includes the sample preparation procedure description used before the test(s). In the future this class will also need to contain much more chemical method and good laboratory practice information. See Table 2.

3.2.1 *sample-amount*—sample amount used to prepare the test material. The unit is milligrams.

3.2.2 *sample-id*—user-assigned identifier of the sample.

3.2.3 *sample-id-comments*—additional comments about the sample identification information that are not specified by any other sample-description data elements.

3.2.4 *sample-injection-volume*—volume of sample injected, with a unit of microliters.

3.2.5 *sample-name*—user-assigned name of the sample.

3.2.6 *sample-type*—indicated whether the sample is a standard, unknown, control, or blank.

3.3 *Definitions for Detection-Method Information Class*—This information class holds the information needed to set up the detection system for an experiment. Data element names assume a multi-channel system. The first implementation applies to a single-channel system only. Table 3 shows only the column headers for a detection method for a single sample.

3.3.1 *detection-method-comments*—users' comments about detector method that is not contained in any other data element.

3.3.2 *detection-method-name*—name of this detection-method actually used. This name is included for archiving and retrieval purposes.

3.3.3 *detection-method-table-name*—name of this detection method table. This name is global to this table. It is included for reference by the sequence information table and other tables.

3.3.4 *detector-maximum-value*—maximum output value of the detector as transformed by the analog-to-digital converter, given in detector-unit. In other words, it is the maximum possible raw data value (which is not necessarily actual maximum value in the raw data array). It is required for scaling data from the sending system to the receiving system.

3.3.5 *detector-minimum-value*—minimum output value of the detector as transformed by the analog-to-digital converter, given in detector-unit. In other words, it is the minimum possible raw data value (which is not necessarily the actual minimum value in the raw data array). It is required for scaling data to the receiving system.

3.3.6 *detector-name*—user-assigned name of the detector used for this method. This should include a description of the detector type, and the manufacturer's model number. This information is needed along with the channel name in order to track data acquisition. For a single-channel system, channel-name is preferred to the detector-name, and should be used in this data element.

3.3.7 *detector-unit*—unit of the raw data. Units may be different for each of the detectors in a multichannel, multiple detector system.

3.3.7.1 *Discussion*—Data Scaling: Data arrays are accompanied by the maximum and minimum values (detector_maximum, detector_minimum, and detector_unit) that are possible. These can be used to scale values and units from one system into values and units for another system. For example, one system may produce raw data from 0 to 100000 counts, and be converted to -100 millivolts to 1.024 volts on another system. This scaling is not done automatically, and must be done by either the sending or receiving system if required.

TABLE 2 Sample-Description Information Class

Date Element Name	Datatype		Category	Required
sample-id-comments	string		C5	. . .
sample-id	string		C1	. . .
sample-name	string		C1	. . .
sample-type	string		C1	. . .
sample-injection-volume	floating-point	C3		. . .
sample-amount	floating-point		C3	. . .

TABLE 3 Detection-Method Information Class

Data Element Name	Datatype	Category	Required
detection-method-table-name	string	C1	. . .
detection-method-comments	string	C1	. . .
detection-method-name	string	C1	. . .
detector-name	string	C1	. . .
detector-maximum-value	floating-point	C1	M1
detector-minimum-value	floating-point	C1	M1
detector-unit	string	C1	M1

3.4 *Definitions for the Raw-Data Information Class*—This is the information actually generated by the data acquisition process. The data are then fed into the peak processing algorithms. This table shows only the column headers for the raw date arrays. Fig. 1 illustrates the exact meaning of the data elements in this information class. See Table 4.

3.4.1 *actual-delay-time*—The time delay between the injection and the start of data acquisition, given in the retention-unit.

3.4.2 *actual-run-time-length*—The actual run time length from start to finish for this raw data array, given in the retention-unit.

3.4.3 *actual-sampling-interval*—The actual sampling interval used for this run, given in the unit of the retention-unit. At this time, it is for a fixed sampling interval.

3.4.4 *autosampler-position*—The position in the autosampler tray. The default datatype for this was chosen to be a string because some companies have concentric rows of sample vials in the sample tray; others may use cartesian coordinates. The format of this is a free-form string, with two substrings, using a period as a delimiter, for example, "coordinate1.coordinate2" or (tray.vial).

3.4.4.1 *Discussion*—Usage of Raw-Data Information: The order of usage for using raw data from this information class is very simple. First check the uniform-sampling-flag to see if it is "Y." If it is, then use only the ordinate-value array for amplitude values, and calculate the abscissa values from point 0.0 onward using the actual-sampling-interval. If the value of uniform-sampling-flag is "N," then use the ordinate-value array for amplitude values and the raw data retention array for abscissa values.

3.4.5 *ordinate-values*—This is a set of values of dimension point-number, containing the ordinate values. This set of values has a unit of detector-unit. This is a required field for datasets containing raw data.

3.4.5.1 *Discussion*—There is no data point at time = 0.0 (or volume = 0.0). The first data point is at the first point after the start of data acquisition.

3.4.6 *point-number*—value of point-number is the dimension of the ordinate-values and (if present) raw-date-retention arrays. It should be set to zero if these arrays are empty.

3.4.7 *raw-data-retention*—This is a set of values of dimension point-number, containing the abscissa value for each raw data ordinate value. This set of values has a unit of retention unit. This is a required field if the uniform sampling flag is "N." *Example:*

raw_data (1) = 12 raw_data_retention (1) = 0.2

raw_data (n-1) = 998760 raw_data_retention (n-1) = 120.1
raw_data (n) = 997650 raw_data_retention (n) = 120.3
raw_data (n+1) = 996320 raw_data_retention (n+1) = 121.5

raw_data (point_number) = 20 raw_data_retention (point_number) = 720.2

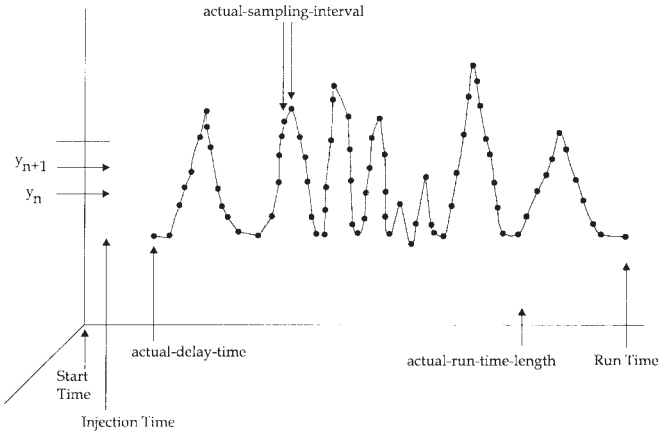

FIG. 1 Raw Data Element Semantics

TABLE 4 Raw-Data Information Class

Data Element Name	Datatype	Category	Required
point-number	dimension	C1	M1
raw-data-table-name	string	C1	. . .
retention-unit	string	C1	M12
actual-run-length	floating-point	C1	M12
actual-sampling-interval	floating-point	C1	M12
actual-delay-time	floating-point	C1	M12
ordinate-values	float-array	C1	M1
uniform-sampling-flags	boolean	C1	M1
raw-data-retention	float-array	C1	M1
autosampler-position	string	C1	. . .

3.4.8 *raw-data-table-name*—name of this table, included for reference by the sequence information table and other tables.

3.4.9 *retention-unit*—unit along the chemical or physical separation dimension axis. All other data elements that reference the separation axis have the same unit.

3.4.9.1 *Discussion*—The developers of the protocol have considered the implications and relative merits of using time versus volume, and is using a "seconds" unit for chromatographic techniques, including Capillary Zone Electrophoresis (CZE) and Size Exclusion Chromatography (SEC). If the user employs CZE or SEC, and wants to use a unit other than seconds, then they should use that as the value of the retention-unit data element.

3.4.9.2 *Discussion*—For liquid and gas chromatography the default unit for the retention axis is time in seconds.

3.4.11 *uniform-sampling-flag*—A value of "N" for this flag indicates that some kind of non-uniform sampling was used. If non-uniform sampling was used, then an array for raw data retention is required. The default value for this is "Y."

3.5 *Definitions for Peak-Processing-Results Information Class*—This is the information generated by the peak processing algorithms. Final processed results may vary from manufacturer to manufacturer. See Table 5.

3.5.1 *baseline-start-line*—starting point of the computed baseline for this peak, given in a unit of retention-unit.

3.5.2 *baseline-start-value*—starting value of the computed baseline for this peak; in a scaled data unit, the unit for this is the same as that of the ordinate variable.

3.5.3 *baseline-stop-time*—ending point of the computed baseline for this peak, given in a unit of retention-unit.

3.5.4 *baseline-stop-value*—starting ending value of the compound baseline for this peak; in a scaled data unit, the unit for this is the same as that of the ordinate variable.

TABLE 5 Peak-Processing-Results Information Class

Data Element Name	Datatype	Category	Required
peak-number	dimension	C2	M2
peak-processing-results-table-name	string	C3	. . .
peak-processing-results-comments	string	C2	. . .
peak-processing-method-name	string	C2	. . .
peak-processing-date-time-stamp	string	C2	. . .
peak-retention-time	floating-point-array	C2	M2
peak-name	string-array	C3	. . .
peak-amount	floating-point-array	C2	M3
peak-amount-unit	string	C2	M3
peak-start-time	floating-point-array	C2	. . .
peak-end-time	floating-point-array	C2	. . .
peak-width	floating-point-array	C2	. . .
peak-area	floating-point-array	C2	M2
peak-area-percent	floating-point-array	C2	. . .
peak-height	floating-point-array	C2	M2
peak-height-percent	floating-point-array	C2	. . .
baseline-start-time	floating-point-array	C2	. . .
baseline-start-value	floating-point-array	C2	. . .
baseline-start-value	floating-point-array	C2	. . .
baseline-stop-time	floating-point-array	C2	. . .
baseline-stop-value	floating-point-array	C2	. . .
peak-start-detection-code	string-array	C2	. . .
peak-stop-detection-code	string-array	C2	. . .
retention-index	floating-point-array	C2	. . .
migration-time	floating-point-array	C2	. . .
peak-asymmetry	floating-point-array	C2	. . .
peak-efficiency	floating-point-array	C2	. . .
mass-on-column	floating-point-array	C2	. . .
manually-reintegrated-peaks	boolean-array	C2	. . .

3.5.5 *manually-reintegrated-peaks*—A boolean flag that indicates if any reported results are based on manual manipulation of baselines or peak start/end times, or both. A value of logical "O" for this flag indicates that the current peak was not manually reintegrated.

3.5.6 *mass-on-column*—A measure of column loading. It is usually reported as the sum of the peak-amount(s). It needs to be determined against known peaks.

3.5.7 *migration-time*—The transit time from the point of injection to the point of detection, given in the retention-unit. This is used in Capillary Zone Electrophoresis (CZE).

3.5.8 *peak-amount*—amount of substance that is determined by peak processing.

3.5.9 *peak-amount-unit*—unit used for the peak amount. This is in a concentration unit, absolute amount in grams, or some other appropriate unit.

3.5.10 *peak-area*—computed area of the peak, given in a scaled data unit of (detection-unit · retention-unit).

3.5.11 *peak-area-percent*—compound area percent of the peak: the summation of all quantified peaks in an analysis should be equal 100.0.

3.5.12 *peak-asymmetry*—The peak asymmetry measured as A_s= B/A, where A and B are the widths for the front and back parts of a peak, commonly measured at 10 % peak height for USP at 5 %.

3.5.13 *peak-efficiency*—Also known as the column theoretical plate number for an individual peak. The peak-efficiency can be expressed as:

$$[(\text{retention}-\text{time} / \text{width}-\text{at}-\text{half}-\text{height})^{\wedge}2] \times 5.54 \text{ or as} \quad (1)$$

$$[(\text{retention}-\text{time} / \text{baseline}-\text{bandwidth})^{\wedge}2] \times 16$$

where baseline-bandwidth is the peak width along the baseline as determined by the intersection points of the tangents drawn to the peak above its points of inflexion. This is usually measured for a reference component, although routinely a representative peak in the chromatogram is chosen and this is reported for it.

3.5.14 *peak-end-time*—ending point of the peak, given in a unit of retention-unit.

3.5.15 *peak-height*—computed height of the peak; in a scaled data unit. The unit for this is the same as that of the raw-data.

3.5.16 *peak-height-percent*—computed height percent of the peak; the summation of all quantified peaks in an analysis should be equal 100.0.

3.5.17 *peak-name*—user-assigned name of the peak. This is an optional field because some peaks may be unknown.

3.5.18 *peak-number*—peak number used to identify a particular peak. The value of peak-number us the dimension of each array in the peak-processing-results information class. This data element should be set to zero if no peak processing results are included in the dataset.

3.5.19 *peak-processing-date-time-stamp*—date-time-stamp for peak processing. Indicates when the data was processed. See dataset-date-time-stamp for ISO standard date time stamp syntax details.

3.5.20 *peak-processing-method-name*—name of the method used for peak processing. This is typically assigned by the end user. It is typically used for archival and retrieval purposes.

3.5.21 *peak-processing-results-comments*—Comments about the peak processing results that are not contained in any other data element in this information class.

3.5.22 *peak-processing-results-table-name*—The name of this table, included to make the dataset self-referential. It is global to this information class.

3.5.23 *peak-retention-time*—The retention time of the peak detected, given in the unit of retention-unit.

3.5.24 *peak-start-detection-code*—Codes that are used to describe how the baselines have actually been drawn. The peak type may be represented by a two-letter code.

The following are examples of peak detection codes:

B = baseline peak, that is, the peak begins and/or ends at the baseline.

P = perpendicular drop, that is, the peak begins and/or ends with a perpendicular drop.

skimmed peak = skimmed peak, that is, the peak is a shoulder peak that is skimmed.

VD = vertical drop, that is, the peak begins or ends at a vertical drop to the skim line (such as between two skimmed peaks).

HP = horizontal projection, that is, the peak baseline starts and/or ends with a horizontal projection.

EX = exponential skim, that is, the peak starts and/or stops with an exponential skim.

PT = pretangent skim, that is, the leading edge of the peak is tangent skimmed.
MN = manual peak, that is, the user forced the baseline at the data level.
FR = forced peak, that is, the user forced the baseline at a user-supplied level.
DF = user forced daughter peak, that is, the user forced a baseline on the side of a fused peak (daughter peak).
LP = lumped peak, that is, peaks values are lumped (added) together until this timed event is turned off.

3.5.25 *peak-start-time*—The starting point of the peak, given in a unit of retention-unit.

3.5.26 *peak-stop-detection-code*—See peak-start-detection-code.

3.5.27 *peak-width*—The calculated width of the peak, given in a unit of retention-unit.

3.5.28 *retention-index*—The retention index, as defined by Kovats, is a measure of relative retention which uses the normal alkanes as a standard reference. Each normal hydrocarbon is assigned a number equal to its carbon number times one hundred. For example, n-pentane and n-decane are assigned indices of 500 to 1000, respectively. Indices are calculated for all other compounds by logarithmic interpolation of adjusted retention times, as shown in the following equation.

$$Ia = 100 \cdot N + 100 \cdot n \cdot (\log tRa - \log tRN / \log tR(N + n) - \log tRN) \quad (2)$$

where:
Ia = is the retention index of the peak
N = is the carbon number of the lower n-alkane
n = is the difference in carbon number of the two n-alkanes that bracket the compound
tRa = is the adjusted retention time of the unknown compound
tRN = is the adjusted retention time of the earlier eluting n-alkane that brackets the unknown
tR(N+n) = is the adjusted retention time of the later eluting n-alkane that brackets the unknown

4. Objectives and Features of the Analytical Data Interchange Protocol

4.1 *Technical Objectives*:

4.1.1 *Standards Development and Systems Selection*—The technical goals have been to develop a protocol for analytical data representation and interchange that meets the following criteria:

4.1.1.1 Easy to use by software developers and end users.

4.1.1.2 Readable by humans using some facile mechanism

4.1.1.3 Open, extensible, and maintainable.

4.1.1.4 Applies to multidimensional data (for hyphenated techniques) as well as two-dimensional data.

4.1.1.5 Independent of any particular communication link, like RS-232, IEEE-488, Local Area Networks, etc.

4.1.1.6 Independent of a particular operating system like DOS, OS/2, UNIX, VMS, MVS, etc.

4.1.1.7 Independent of any particular vendor, and acceptable and usable by all.

4.1.1.8 Coexists with, and does not negate, other standards.

4.1.1.9 Designed for the long-term and implemented for use in the short-term,

4.1.1.10 Works well for chromatography and does not preclude extensions to other analytical technique families.

4.1.2 *Data Integrity Across Heterogeneous Systems*—The current implementation specifies a mechanism with particular directionality for data transfer integrity. The protocol has unidirectional data integrity for data transfers between heterogeneous systems. This is because source systems and target systems are made by different manufacturers, or if the systems are from the same manufacturers, they may use different hardware or algorithms. An example would be data transfer from Vendor A's data system running on a DOS-based personal computer to Vendor Z's LIMS running on a Unix-based minicomputer; another would be transfer between chromatographic data systems made by different manufacturers.

4.1.2.1 If the receiving system has algorithms that assume a different analog-to-digital (ADC) converter word length from the sending system, and it calculates results based on its own, different data precision and accuracy, then the accuracy and precision of the original data is going to be maintained. For example, if the sending system has an algorithm that assumes a 24-bit internal representation, and the receiving system has an algorithm that assumes a 20-bit internal representation, one may lose data accuracy and precision. If calculations are done by the receiving system, and the data are then sent back to the source system for their calculations, data integrity may not be maintained. Thus, there is an inherent directionality to data transfer given by different algorithms and different hardware systems.

4.1.2.2 The protocol for chromatographic data can be used for data round-trips relative to the source system, for example, from the source system to an archive and then back to the source system again. Such round-trip data transfers will maintain data integrity as long as there was no calculation or alteration of the data during transfer that would alter its accuracy or precision.

4.1.2.3 Thus, the protocol is bidirectional for homogeneous source-system round trips and unidirectional for heterogeneous source-to-target transfers.

4.1.2.4 The first implementation allows transfer of chromatographic raw data and final results. Plotting and requantitation of raw data on other vendor's data system (for comparison purposes), and transfer of final results to information systems (such as a LIMS) are possible in the first implementation

4.1.3 *Algorithmic Issues*—Algorithmic issues are not addressed at all by this specification. Users cannot expect to get the same exact processed results from systems that use completely different algorithms.

4.1.4 *Absolute Scaling of Raw Data*—Absolute scaling of raw data across different manufacturer's systems is not possible at this time, due to the lack of general-purpose algorithms that can convert and scale data of different internal representations, from different data acquisition systems, and different computer hardware systems.

4.1.5 *Requirements*— The protocol in this specification does not yet specify all elements needed to meet documentation quality data requirements (Good Laboratory Practices or ISO 9000).

4.2 *Technical Features of the AIA Protocol*:

4.2.1 *Separation of Concept from Implementation*—There is a clean separation of the protocol into contents (the data definitions within a data model) and container (the data interchange system). This is important because it effectively decouples concept from implementation. Computer technology is changing much more rapidly than analytical data definitions, which are stabilizing for the maturing analytical instrument industry. Producing an accurate analytical information model and having well-defined definitions for data elements within that model actually have higher long-term significance than any particular data interchange system technology.

4.2.2 *General Technical Features*—Two general technical features stand out:

4.2.2.1 *Analytical Information Categories*—a convenience for simplifying the work of developing analytical data specifications. These five categories were chosen based on three practical considerations: (*1*) which data is of interest to transfer most routinely, (*2*) which can be standardized most easily in the short-term, and (*3*) which can be standardized in the long term. The analytical information categories are explained later in Section 5.

4.2.2.2 *The Data Interchange System*—the container used to communicate data between applications, in a way that is independent of both computer platforms and end-user applications. The system has software routines that are used to read, write, and manipulate data in analytical datasets. It has a data access interface, called an Application Programming Interface (API).

4.2.2.3 The data interchange system that most closely fits the scientific and software engineering requirements for a public-domain data interchange software system is the NetCDF (network Common Data Form) system. The Unidata Corporation, which supports the National Center for Atmospheric Research, is the source of NetCDF. NetCDF is copyrighted by the Unidata Corporation. The Protocol used the NetCDF system for the implementation of its Protocol. The engineering "beta" tests that prove applicability of NetCDF for analytical data applications have been completed. An overview of NetCDF is given in Guide E 1948.

5. Analytical Information Categories

5.1 Data and information usage varies widely in complexity and completeness. Information is therefore sorted into logical categories, called the Analytical Information Categories. These categories serve two very useful purposes.

5.1.1 First, the categories sort analytical information into convenient sets to allow more rapid standardization. This has made it easier for implementors to produce working demonstrations, without the burden and complexity of the hundreds of data elements contained in a full dataset for any given analytical technique.

5.1.2 Second, the categories accommodate different organizations' usage of information more easily. Some organizations may only want to transfer raw data among data systems. Others may want to transfer information to a LIMS or other database systems. Still others may want to build databases of chemical methods, instrument methods, or data processing methods. The first version of the protocol is for a single sample injection, not for sequences of samples.

5.1.3 The information contained in this specification represents the greatest common subset of information end-user and vendor requirements available at this time.

5.2 *Category 1: Raw Data Only*—Category 1 is used for transferring raw data. It includes raw data, units, and relative data scaling information. This will allow accurate replotting of the chromatogram and/or reprocessing. Category 1 also contains administrative information needed to locate the original chemical and data processing methods used with this dataset.

5.3 *Category 2: Final Results*—All post-quantitation calculated results are included. This information category includes the amounts and identities (if determinable) of each component in a sample. Final sample peak processing results, component identities, sample component amounts, and other derived quantities of interest to the analyst are included in Category 2 datasets. Quantitation decisions are included here as comments to aid the analyst in determining how the results were calculated.

5.3.1 Category 2 datasets can be used to transfer data to database management systems, such as a LIMS, research database, or sample tracking systems. It can also be used to transfer data to data analysis packages, spreadsheets, visualization packages, or other software packages.

5.4 *Category 3: Full Data Processing Method*—Quantitation decisions and data processing methods are transferred in this category. Quantitatively correct data/information transfer is achieved by the category for all parameters necessary to do peak detection, measurement, and response factor calculation, and calibration for a sequence of related sample runs. This applies to both samples and reference standards. Sample quantitation results are not included here; those are in Category 2. Peak processing method parameters, response factor calculation and other calibration method parameters required to quantitate sample component peaks are included in Category 3.

5.5 *Category 4: Full Chemical Method*—All chemical method information needed to repeat the experiment under exactly the same chemical conditions is included in this category.

5.6 *Category 5: Good Laboratory Practice Information*—Any additional information required to satisfy Good Laboratory Practices or ISO-9000 requirements are included in this category. This category generally deals with capturing product, process, and documentation quality information needed for validation.

6. Keywords

6.1 analytical information categories; andi; chromatographic; detection method; information class; ISO; NetCDF; peak-processing-results; raw data; sample description