Designation: E 1947 – 98

# 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:*

E 1948 Guide for Analytical Data Interchange Protocol for Chromatographic Data[3]

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.

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

[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 10, 1998. Published August 1998.

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

[3] *Annual Book of ASTM Standards*, Vol 14.01.

[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 | . . . |

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.

**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 | . . . |

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.

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

**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 |

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:*

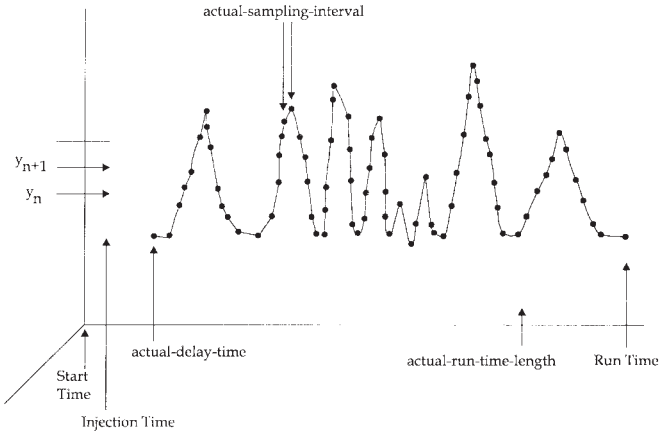


**FIG. 1 Raw Data Element Semantics**

**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 | . . . |

| | |
|---|---|
| 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 |

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.

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