CRSP® CENTER FOR RESEARCH IN SECURITY PRICES
An Affiliate of the University of Chicago

CRSP METADATA GUIDE

STOCK & INDEXES FLAT FILE FORMAT 2.0 (CIZ)

DISCLAIMER

CENTER FOR RESEARCH IN SECURITY PRICES, LLC

An Affiliate of the University of Chicago Booth School of Business

105 West Adams, Suite 1700
Chicago, IL 60603

Phone: 312.263.6400
Fax: 312.263.6430

Email: support@crsp.org

Website: www.crsp.org

Last Modified October 20, 2023

OVERVIEW

With the CRSP Stock and Index Flat File Format 2.0 (aka CIZ), CRSP has included ten metadata files. Three schema files contain information about the files, items, and columns of all the files and can help speed the importing of the ASCII files.

Two files provide easy access to the flag values, descriptions, and definitions used throughout the file rather than having them only in online PDF documents.

Two coverage files contain the results of data profiling done by CRSP that are intended to help provide a more three-dimensional description of the data than a text-only explanation provides.

Two calendar files are intended to improve transparency related to exchange holidays, closures, and ease of use by supplementing and complementing date arithmetic functions with pre-calculated information about the CRSP periods.

The tenth metadata file is helpful for those familiar with the Flat File Format 1.0 (SIZ) files or the CRSPAccess files by providing a mapping between the previous item names and the new items names. Details on that file are found here.

FILE, ITEM, AND COLUMN (SCHEMA) INFORMATION

There are three metadata files that describe the Flat File Format 2.0 (CIZ) files, items, and columns (i.e. schema). These files are MetaFileInfo, MetaItemInfo, and MetaColumnInfo. There are three primary uses for these files:

- For those using the ASCII format files, these files provide the information necessary to create an appropriate database schema or data structure to load the data. For example, is a column an integer, date, string, floating point number, or character string. If a string, what is its maximum width and is it a fixed width field or a variable width field. It should be possible to use these metadata files as input to reduce the time needed for "create table" scripts. In addition, the MetaFileInfo includes information about the columns used to sort the file and uniquely identify a row.
- While the CRSP Stock & Indexes Databases: Flat File Layout 2.0 Guide provides much of the same information, these files allow a user to do more complex searches of the item names, descriptions, and definitions in their tool of choice (SAS, 'R', SQL Server, etc.) to more accurately and quickly identify the columns of interest from more than 500 columns among all the files. These files can also be used to automate report headers or variable labels.
- The Item Category and Item Class along with some columns in the MetaColumnInfo provide useful information about the content of a column to better determine how to use or output the data.

FLAG FILES

CRSP provides nearly 300 distinct items in more than 400 columns where the value is a flag. The MetaFlagInfo and the MetaFlagType files provide the supporting information about these flags. When the ItemFlagType column in the MetaItemInfo file is not NULL, it contains a Flag Type.

These Flag Types can be looked up in MetaFlagType for information about the type and in MetaFlagInfo to find all the valid Flag Values for this Flag Type. These flags help describe or qualify the entity or value and may be used to subset or group the data in queries.

Some of these flags are self-explanatory. For example, a YES/NO flag with values of 'Y' and 'N' or a data frequency flag of 'A' for Annual, 'Q' for quarterly, 'M' for monthly, and 'D' for daily. Others are less mnemonic and intuitive, and all but the most frequent user of the data will need to look up the flag values to ensure they understand the data. There are three uses of the files:

- While the Flag Value Table on the CRSP website provides much of the same information, these two files allow a user to do more complex searches of the flag values, descriptions, and definitions in their tool of choice (SAS, 'R', SQL Server, etc.) to more accurately and quickly identify the flags of interest out of the nearly 1,000 unique FlagType/FlagValue combinations than is possible from PDF documents or website tables.

- These files may also be used to include flag descriptions into report output to make it more understandable to a reader not family with CRSP flags.
- For database administrators or application creators, the Flag Types and/or Flag Values and their corresponding descriptions and definitions can be used to populate selection options (e.g., drop-down menus) in query tools.

COVERAGE FILES

Two metadata files in the Flat File Format 2.0 (CIZ) provide coverage information. These files are MetaColumnCoverage and MetaFlagCoverage. The coverage file values are calculated from the 1925 US Stock and Index File (CIZ) and for some items and flags do not represent what is found in the subset files (e.g. C6Z – 1962 US Stock File). The stock and index data files may be used without reviewing the new coverage files, especially for those already familiar with CRSP data; however, the coverage files are being introduced to improve ease-of-use and save users time. There are three primary uses for these files:

- Help determine which column(s) might be most beneficial before running potentially time-consuming queries on the data, especially the very large daily data files;
- Help determine which, if any, flag values may impact the design of a query;
- Profiling information that could be useful for improving query performance or reducing storage.

A row exists in MetaColumnCoverage for any column with ColCoverageFlg in the MetaColumnInfo file set to 'Y'. The MetaColumnCoverage file includes more than 100 columns and almost all non-key, non-flag, and non-metadata columns.

A row exists in MetaFlagCoverage for any FlagType/FlagValue combination in the MetaFlagInfo file where FlagCoverageFlg = 'Y'. The MetaFlagCoverage file includes over 300 FlagType/FlagValue combinations and is all non-metadata items that are Item Category of FLAG.

COLUMN COVERAGE AND USEFULNESS OF COLUMNS

There are more than 500 columns in the Flat File Format 2.0 (CIZ) files; some have very similar names and descriptions but very different characteristics. The MetaColumnCoverage file can provide easy-to-access information to determine which column(s) are most appropriate for the intended use. For example, DlyPrc has 98% non-missing values and at least one non-missing value for 100% of the securities (PERMNOs) in the file, but DlyClose has only 81% non-missing values and at least one non-missing value for only 88% of the securities, and DlyOpen only has 60% non-missing and at least one-non-missing value for only 77% of the securities. If a closing price versus a bid-ask average is not an important distinction to study, then DlyPrc has significantly more (about 17,000,000) non-missing values. A study comparing opening and closing prices will be limited by the availability of the DlyOpen.

Similarly, the number of trades (DlyNumTrd) might be helpful for a study but knowing that its first non-missing data is on 11/1/1982 and has at least one non-missing value for only 48% of the securities could impact the study design.

FLAG COVERAGE AND DATA RESTRICTIONS OR GROUPINGS

There are more than 300 distinct FlagType/FlagValue combinations. Some combinations indicate outliers, and some queries will often exclude them because they are not relevant or material to the study. Excluding those rows will keep the data cleaner and simplify the coding and analysis.

Conversely, sometimes those very same outlier combinations are of particular interest to a study and knowing the magnitude of those outliers can determine if enough data exists for robust analysis. For example, SecuritySubType (FlagType='S2') has rows for its five values indicating about 31,000 Common (starting in 1925), 4,000 Exchange Traded Funds (starting in 2002), 1,300 Closed-End Funds (starting in 1962), 81 Americus Trusts (only from 1983 to 1992), and 54 Exchange Trade Vehicles (starting in 1987). Some studies may restrict data to Common only. In contrast, others might want to compare the performance broken down by SecuritySubType, while others may view SecuritySubType as an unimportant distinction and ignore it.

PROFILING INFORMATION FOR IMPROVING QUERY PERFORMANCE OR REDUCE STORAGE

While the coverage files are not intended to replace advanced data profiling needed for sophisticated database tuning, the coverage files may be helpful as a starting point for database administrators trying to improve query performance or reduce storage. For example, more sparsely populated columns could be identified and, if appropriate, be moved from commonly used tables to supplemental tables to reduce the size of the widely used table and speed access but not significantly reduce the table's usefulness. Flag values that are commonly used and that effectively split up the file could be indexes to speed queries access that table.

CALENDAR FILES

Two metadata calendar files, MetaExchangeCalendar and MetaCalendarPeriod, are included in the Flat File Format 2.0 (CIZ) files. These files provide information that complements and supplements the date functionality of SAS, 'R', and SQL-based databases.

MetaExchangeCalendar – TRADING DAY, HOLIDAY, AND WEEKEND FLAGS

The MetaExchangeCalendar file contains every day from the start of the CRSP data file from December 31, 1925, to the last trading of the files. In addition, it provides flags for whether the day is a trading day, a holiday, or a weekend. It also has a code that combines the flags and details the reasons for the holiday.

While trading days in recent years follow a regular pattern, over the nearly one hundred years CRSP tracks, there is much more variability than most remember, including Saturday trading, Thanksgiving not always being the fourth Thursday of November, the standardization of observing several holidays on Monday, the recent addition of Juneteenth, etc. In addition to the planned changes to trading days, there were unexpected closures, including the multi-day closure in September 2001, the single-day closures for hurricanes, presidential funerals, bank runs, and other unforseen events. While most studies do not need to control for weekends and holidays, the flags in the MetaExchangeCalendar are available to provide transparency. They can be a time saver in identifying an external cause of anomalous days.

MetaCalendarPeriod – INFORMATION FOR DAILY, MONTHLY, QUARTERLY & ANNUAL PERIODS

The MetaCalendarPeriod file provides information about the Daily (is daily going to be added in, if not do we document it eventual and footnote the initial deficiency or change the doc and then change it again after it is fixed), Monthly, Quarterly, and Annual Calendar periods that can be used to complement and supplement the default date arithmetic functions.

The file contains the calendar and CRSP trading day start and end dates for a period. For example, the annual period for 1994 has the expected calendar range of 1/1/1994 to 12/31/1994, but its CRSP trading range is from 1/3/1994 to 12/30/1994. In addition, it contains the next and previous periods, which for years is simple +1 or -1 (e.g., next and previous year for 1994 is 1995 and 1993 respectively), but for months and quarters are not as simple to calculate. This allows an easy join.

It also contains deformalized data such as CRSPPeriodPrevCRSPEndDt and has counts of both the number of calendar days in the period and the number of trading days and a CalPeriodNbr so that it is easy to calculate the number of periods between two dates. The information in the MetaCalendarPeriod file is not needed for most queries. It can be derived from other CRSP data, but having it pre-calculated and available can be a great time saver when needed.

SIZ TO CIZ MAPPING INFORMATION

The MetaSIZtoCIZ file is designed to provide detailed information about how the CRSP Flat File Format 1.0 (SIZ format) maps to the CRSP Flat File Format 2.0 (CIZ format). The complete list of columns in this file are in the CRSP Stock & Indexes Databases: Flat File Layout 2.0 Guide. This document section summarizes the data's potential uses in the MetaSIZtoCIZ file.

The MetaSIZtoCIZ file contains more detail than what is available in the Cross Reference Guide, but serves a similar purpose. This file contains multiple types of information, including primary and supplemental join information and column mapping details.

The SIZColPosition (SIZ Column Position number) and SIZColMapSeq (SIZ Column Mapping Sequence number) are the columns that provide the information needed to determine which type of information the row in the MetaSIZtoCIZ file contains, and along with the SIZFileName make up the sort order and natural key of this file. The SIZtoCIZType (SIZ to CIZ mapping Type) and SIZtoCIZSubType (SIZ to CIZ mapping Sub-Type) columns then provide detailed information on the map specific to that row.

There are three primary uses for this file:

- A more detailed and filterable way to look up an existing SIZ column to determine the CIZ column(s) that are the closest match to a SIZ column than what is available in the higher level Cross Reference Guide. The file is in SIZFileName, SIZColPosition order to efficiently facilitate the review of all the columns for one SIZ file. While many SIZ columns map to one and only one CIZ column, SIZ contain several overloaded values that have been split into multiple CIZ columns for a more modern database design.
- Information about the relationships of rows in the SIZ file compared to the rows in the corresponding CIZ file. While most of the largest files SIZ map one-to-one with CIZ files and substituting one for the other should require no change in code logic, there are others where there are differences in implementations that might require some modification of code.
- Information about some new columns or files in CIZ and what is its closest SIZ match.

The remainder of this section describes the four columns (SIZColPosition, SIZColMapSeq, SIZtoCIZType, SIZtoCIZSubType) that are most important in understanding the contents of the MetaSIZtoCIZ mapping file. Appendix XXXX has a complete layout and column description of this file.

SIZColPosition – SIZ COLUMN POSITION NUMBER

The SIZColPosition (SIZ Column Position number) is used along with SIZFileName to sort the file. Therefore, the rows will match the layout of the SIZ file, rather than an alphabetical listing of items, and hopefully make it easy to review the changes on a file-by-file basis.

Most commonly, the SIZ Column Position Number is between 1 and 99 and is simply the column position number within the SIZ file; 1 indicates the first column, 2 the second column, etc. When the SIZColumnPosition number equals zero or exceeds 900, the row contains join information or added value, respectively.

COLUMN MAPPING – SIZColPositioN BETWEEN 1 AND 99 (CURRENT MAXIMUM 31)

These rows are the most common and most straightforward use of the file. For example, the 4th column of the SFZ_AGG_MTH file (MTHPRC) maps to the MthPrc column in the StkMthSecurityData file. Therefore, for this row, the corresponding mapping type and sub type (see sections below) indicate that it is an exact match within machine precision. However, some columns have more complicated mappings, including some columns having multiple mappings (see SIZColMapSeq section below).

JOIN MAPPING – SIZColPosition EQUAL TO 0

These rows indicate how the key value(s) in the SIZ file align with the key value(s) in the corresponding CIZ file and help determine how to join or merge the files. For these rows, the SIZItemName has two valid values with more details in the

SIZItemDesc and CIZFileName, CIZItemDesc

- Key – the join will be a single column key (e.g. KYPERMNO to PERMNO)
- KeyCombo – the join will be a combination of columns (e.g. KYPERMNO, YYYY to PERMNO, YYYY)

JOIN INFORMATION – SIZColPosition EQUAL TO 0

These rows indicate how the key value(s) in the SIZ file align with the key value(s) in the corresponding CIZ file and help determine how to join or merge the files. For these rows, the SIZItemName has two valid values with more details in the SIZItemDesc and CIZFileName, CIZItemDesc

- Key – the join will be a single column key (e.g. KYPERMNO to PERMNO)
- KeyCombo – the join will be a combination of columns (e.g. KYPERMNO, YYYY to PERMNO, YYYY)

The SIZtoCIZType (see page 8) provides information on what type of relationship the files have (one-to-one, ZeroOne-to-One, etc.). The SIZtoCIZSubType (see page 8) includes additional information to help understand the relationship.

CIZ Added Value FOR EASE OF USE – SIZColPosition GREATER THAN 900

When ColPosition > 900, the SIZtoCIZType is either "Denormalized" or "Documentation". The "Denormalized" columns (COMNAM, INDFAM, and PORTNUM) are columns that exist in SIZ but had to be found in another SIZ file.

These rows indicate columns that were included in the corresponding CIZ file for ease of use and to reduce the need to join multiple SIZ files.

Similarly, the rows where SIZtoCIZType is "Documentation" type are for columns that were implicit in the SIZ file because the information was contained in the PDF document. For example, IndexStatType (Index Statistic Type) – the Index Calculation section of the PDF Data Description had the relationships between INDNOs and Index Statistic Type. Still, that information has been moved to the CIZ file to reduce the need to look up information in the PDF guides to improve ease of use. These rows are also helpful for determining the closest equivalent in SIZ to a new CIZ column.

SIZColMapSeq – SIZ COLUMN MAPPING SEQUENCE NUMBER

The SIZCoMapSeq (SIZ Column Mapping Sequence number) handles cases where a single SIZ column is split into two or more CIZ columns, or a single SIZ column has two or more different mappings to the CIZ or both situations. It also indicates multiple join possibilities: the lower the sequence number, the more common the mapping usage.

As a general rule, if in doubt, the row with SIZColMapSeq=1 is the best default mapping, but where there are multiple rows, it is best to review the additional rows to ensure that the best default mapping is, in fact, the best mapping for a specific application.

SINGLE COLUMN SPLIT TO TWO OR MORE CIZ COLUMNS

For example, SFZ_DEL file's 3rd column DLSTCD is three-digit code numeric code that is split into four CIZ fields: DelActionType; DelStatusType; DelReasonType; and DelPaymentType. See the Numeric Code split section in the Summary of Changes document for more information about these types of splits.

SINGLE COLUMN TO TWO OR MORE MAPPINGS

For instance, the SFZ_DP_DLY file's 4th column RET is a very commonly used item, but there are two mappings. The SIZColMapSeq = 1 has the most direct mapping to CIZ column DlyRet in the StkDlySecurityPrimaryData file, but it also maps to DlyRet in StkDlySecurityData (SIZColMapSeq=2). These two versions of DlyRet are identical and which to use depends on what other data are needed. StkDlySecurityPrimaryData has fewer columns, providing faster access, and is most analogous to SFZ_DP_DLY. However, if data from StkDlySecurityData is needed, it is possible to get at it using that only one file with no need to merge or join.

BOTH SITUATIONS

For instance, the SFZ_DP_DLY file's 3rd column PRC is a very commonly used item, with seven mappings. Like DlyRet, DlyPrc is available in two different files: StkDlySecurityPrimaryData (SIZColMapSeq=1) and in StkDlySecurityData (SIZColMapSeq=2). Also, the convention is SIZ for PRC was to set it to a negative value to indicate a bid-ask average and positive when it was a closing trade. This forced the use of an absolute value function before calculating.

In CIZ, the DlyPrc is the absolute value of PRC, and a new field, DlyPrcFlg (SIZColdSeqMap=4 and 5), was added to more easily allow users to differentiate and filter among a closing trade, a bid-ask average, and a missing value if desired. Still, the absolute value function is no longer necessary.

The SFZ_DP_DLY PRC column is also the basis for the new DlyClose (SIZColSeqMap=3, DlyPrevPrc (SIZColSeqMap=6), and DlyPrevPrcFlg (SIZColSeqMap=7), columns.

MULTIPLE JOIN POSSIBILITIES

There are a few broad cases where multiple joins exist. One case is when there are two or more natural keys because of redundant keys. For example, SFZ_AGG_MTH could be joined to StkMthSecurityData on either KYPERMNO, YYYYMM to PERMNO to YYYYMM (SIZColSeqMap=1) or on KYPERMNO, MCALDT to PERMNO to MthCalDt (SIZColSeqMap=2). The YYYYMM and MthCalDt are redundant, and which to use is dependent on the implementation of indexing, if any, in the environment, and the key columns of other data, if any, that will also be looked at. A second case is where different tables exist. For example, the SFZ_HDR file joins on KYPERMNO to CIZ's StkSecurityInfoHdr's PERMNO (SIZColSeqMap=1) but also joins PERMCO to the new CIZ StkIssuerInfoHdr's PERMCO (SIZColSeqMap=2).

SIZtoCIZType – SIZ TO CIZ MAPPING TYPE

The SIZtoCIZType (SIZ to CIZ mapping Type) is flag type 'MT' and values for this column can be found in MetaFlagInfo. The row type (see SIZColPosition section above) indicates the type, and the SIZtoCIZTypes are specific to the row type.

SIZtoCIZSubType – SIZ TO CIZ MAPPING SUB-TYPE

The SIZtoCIZSubType (SIZ to CIZ mapping Sub-Type) is flag type 'MU' and values for this column can be found in MetaFlagInfo. The SIZtoCIZSubType, as its name implies, provides additional information about the SIZtoCIZType and should be looked at in combination.

APPENDIX

MetaSIZtoCIZ – SIZ TO CIZ COLUMN MAPPING

File includes information about the SIZ to CIZ Column Mapping and useful to translate from SIZ columns to CIZ columns and conventions.

Item Name	Item Description	Item Definition
SIZFileName	SIZ File Name	Name of the SIZ File
SIZColPosition	SIZ Column Position	Column Position of the item within the SIZ file
SIZColMapSeq	SIZ Column Mapping Sequence	Sequence number to distinguish when then is a one to many mapping between an SIZ Item and a CIZ Item
SIZItemName	SIZ Item Name	Name of the SIZ Item
SIZItemDesc	SIZ Item Description	Short Description of the SIZ Item
CIZFileName	CIZ File Name	Name of the CIZ File, if a mapping is available
CIZItemName	CIZ Item Name	Name of the CIZ Item, if a mapping is available
CIZItemDesc	CIZ Item Description	Short Description of the CIZ Item, if a mapping is available
SIZMappingType	SIZ to CIZ Mapping Type	Mapping Type provides information on how the SIZ column could be mapped to the CIZ column(s), if applicable
SIZMappingSubType	SIZ to CIZ Mapping Sub Type	Mapping Sub Type provides additional information on how the SIZ column could be mapped to the CIZ column(s), if applicable
CIZColumnKey	CIZ Column Key	CIZ Column Key is a unique key for the Metadata Column Information File
CIZFileKey	CIZ File Key	CIZ File Key is a unique key for the Metadata File Information File
CIZItemKey	CIZ Item Key	CIZ Item Key is a unique key for the Metadata Item Information File
SIZColumnMappingKey	SIZ to CIZ Mapping Key	SIZ to CIZ Mapping Key is a unique surrogate integer key to the MetaSIZtoCIZ file