User's Guide to Thomson Reuters

Mutual Fund and Investment Company

Common Stock Holdings Databases on WRDS

Wharton Research Data Services

July 2008

I. Overview

Stock holdings by financial institutions are widely studied by finance scholars in the context of measuring investment performance, studying the rising popularity of mutual funds, characterizing trading patterns of large institutional investors, and even investigating corporate governance issues. The most commonly used databases for these studies have been the Thomson Financial sets that are also known as CDA/Spectrum S12 ('S-one-two') and S34 ('S-three- four'), with the former covering individual mutual funds and the latter covering entire investment companies. The investment companies, which include banks, insurance companies, parents of mutual funds, pension funds, university endowments, and numerous other types of professional investment advisors, are often called 13f institutions, referring to the form that they are required to file with the SEC every quarter.

The Thomson sets are available on WRDS as part of the Thomson Financial Network (TFN) group with holdings data starting in the first quarter of 1980. Both the mutual fund and institutional sets provide detail on holdings of US equities, and a limited set of other securities, including foreign stocks, and are based on documents filed by registered investment companies and professional money managers.[1]

This document serves as an explanation of the form in which the Thomson sets are organized and archived on WRDS, describes the various data items that are in the two sets, and includes additional information to help researchers understand the level and accuracy of the holdings coverage.[2] Problems and limitations with certain aspects of the data are also noted and discussed.

It is useful to know that the Mutual Fund (S12) and Investment Company (S34, or 13f

[1] For detail on filing rules, see the SEC website. For an example copy of the 13f form, see http://www.sec.gov/about/forms/secformsalpha.htm. There are no blank, paper based N-30D forms for the annual or semi-annual reports that an investment company (such as a mutual fund) mails to shareholders pursuant to Rule 30e-1 of the Investment Company Act. (The SEC does not control or maintain the site that includes the text of Rule 30e-1.) Normally, N-30D is accepted as an electronic submission to the SEC through the EDGAR system.

Institutions) sets are related and share a similar structure, but they differ in their original sources and in their level of coverage.[2] The basic relationship between the sets comes from the fact that almost every fund in the S12 set has a manager in the S34 set, and the latter provides aggregated totals for the holdings of all funds under the manger's control (i.e. a fund family). For example, Fidelity (MGRNO=27800) reports as a single entity and aggregates the holdings of all funds and trusts that it manages into its quarterly 13f filings. Fidelity also reports holdings of individual funds such as Magellan (FUNDNO=21858), its largest equity fund, both in fund prospectuses and official SEC filings.

In other words, details on the holdings of Magellan and other Fidelity funds are available in the S12 sets, while the aggregate for all Fidelity funds is in the S34 set. The statutory requirements for reporting mutual fund holdings are only semi-annual, however, and unfortunately the data has reporting gaps for many mutual funds. In addition, the 13f universe is much larger than mutual fund managers, including for example, the investment portfolios of trust departments at banks and other types of institutions that are similar to mutual fund managers, but technically not the same as a mutual fund.

The holdings are identified by CUSIP and are generally equities, but are not necessarily the entire equity holdings of the manager or fund. The potential exclusions include: small holdings (typically under 10,000 shares or $200,000), cases where there may be confidentiality issues, reported holdings that could not be matched to a master security file, and cases where two or more managers share control (since the SEC requires only one manager in such a case to include the holdings information in their 13f report).

Section II provides an introduction to data sources, coverage, and describes the original Thomson data files or tables. Section III describes the manner in which WRDS joins these tables using data items such as FUNDNO, FDATE, RDATE, and CUSIP and creates the all-inclusive S12 and S34 databases. Data anomalies and issues related to missing and stale data that could create problems for users are discussed in Section IV.

[2] Thomson Financial has not produced a formal manual for the Thomson sets, but has provided various supporting documents that were used to help produce this document.

II. Data Item Summary

In Table 1 below is a list of each variable or data item in the S12 and S34 data sets, along with a short descriptions, the table 'types' or source file (denoted as S12n and S34n), and additional useful information.

Table 1: Data Items

Variable Name	Description	TFN Source Files	Data Type	Start Date
ASSETS	End of Qtr Assets (x10000)	S121	num	1980
CHANGE	Net Change in Shares Since Prior Report	S124, S344	num	1980
COUNTRY	Country	S121, S341	char	1999 *
CUSIP	Cusip (Security ID) 8 character version	S122, S123, S124, S342, S343, S344	char	1980
EXCHCD	Exchange Code	S122, S342	char	1999 *
FDATE	File Date	All Files	date	1980
FUNDNAME	Fund Name	S121	char	1980
FUNDNO	Fund Number	S121, S123, S124	num	1980
INDCODE	Industry Code	S122, S342	num	1999 *
IOC	Investment Objective Code	S121	num	1980
MGRCOAB	Management Company Abbreviation	S121	char	1980
MGRNAME	Manager Name	S341	char	1980
MGRNO	Manager Number	S341	num	1980
NOVSHRS	No Voting Authority Shares Held	S343	num	1999 *
PRC	End of Qtr Share Price	S122, S342	num	1980
PRDATE	Prior Report Date	S121, S341	date	1999 *
RDATE	Report Date	S121, S341	date	1980
SHVSHRS	Shared Voting Authority Shares Held	S343	num	1999 *
SHARES	Shares Held at End of Qtr	S123, S343	num	1980
SHROUT1	End of Qtr Shares Outstanding in Millions	S122, S342	num	1980
SHROUT2	End of Qtr Shares Outstanding in 1000s	S122, S342	num	1980
SOVSHRS	Sole Voting Authority Shares Held	S343	num	1999 *
SPRDATE	Derived Stock Price Date	S12, S34	date	1980
STKCD	Stock Class Code	S122, S342	char	1999 *
STKCDESC	Stock Class Description	S122	char	1999 *
STKNAME	Stock Name	S122, S342	char	1980
TICKER	Ticker Symbol	S122, S342	char	1980
TICKER2	Extended Company Ticker Symbol	S122, S342	char	1994 *
TYPE1	Manager Type Code	S341	num	1980

* A set of data items were added to the data files in 1999 (i.e., not provided in the original set of CDA data files). The 1999 additions are: INDCODE MGRCOAB COUNTRY PRDATE

EXCHCD STKCD NOVSHRS SHVSHRS SOVSHRS and SHROUT2. TICKER2 was first included in 1994 data for the S34 set, but became more consistently available with the changes in 1999. EXCHCD and STKCD were placed in 1993:4 (for S12 set) and 1994:1 (for S34 set) files but are not are consistently reported until 1999. STKDESC is only reported for the S12 set in 1999. Other items, such as MGRCOAB and IOC not available for all FUNDNOs and MGRNOs in all years, as well.

All items above are both in distinct table 'types' that cover different aspects of mutual funds or investment companies and their holdings and in the all inclusive S12 and S34 sets. The latter two sets were constructed to make the data easier to use. Still the individual sets are available to the researcher via web query or the WRDS server.[3]

Additional Notes:
1. Both the FUNDNO and MGRNO identifiers are reused in the data. In other words, FUNDNO or MGRNO do not provide unique and permanent identifiers for every fund or manager. A gap of more than 1 year in the RDATE for the same FUNDNO or MGRNO typically reflects a different and unrelated fund or manager. In other cases, a fund or manager may be reassigned a different FUNDNO, usually when a name changed occurred. Therefore a terminal RDATE for a fund or manager does not necessarily signify that the fund or manager is no longer operating. Most FUNDNAME and MGRNAME changes are not associated with a different FUNDNO or MGRNO, but instead reflect an official name change or a different manner of abbreviating the same name (i.e ALLEN B. SMITH INVESTMENT MANAGEMENT -> AB SMITH INVT MGMT CO).

2. The PERMKEY identifier from Managers in the S34 set is not consistently populated starting in 2002, and Thomson treats this item as an obsolete, legacy identifier.

3. The TYPECODE variable in the S34 set indentifies the type of institution (banks, insurance companies, investment companies, independent advisors). This variable should not be confused with the TFN file 'types'. The TYPECODE variable is not reliable from 1998 and beyond; it reflects a mapping error that occurred when TFN integrated data from the former Technimetrics. Many of these institutions were and are still improperly classified as TYPE=5 (endowments and others). For example, in the last quarter of 1998, the number of investment companies drops to 0. TFN regrets that the problem occurred but they have no plans to fix the problem.

4. The PRDATE item is defined as 'Prior Report Date' with a correspondence to RDATE in the same type 1 source files, but this item may actually be based on the FDATE definition, i.e. the prior vintage date. This item should be used with caution.

5. The stock price series are generally comparable to CRSP data, but in some cases are stale,

[3] These files were not actually cut or produced on the FDATE, but instead FDATE represents the quarter (last day of quarter) for which the data items were generally available for public information such as stock prices, and for holdings, theoretically available through fund or investment company records.

representing the closing price of the prior quarter, especially in 1999 and 2000 data.

6. Data coverage in March 2000 for the S34 set is different from the rest of the data. TFN reports that lost data files for March 2000 required a restoration from backup at another business unit (not CDA/Spectrum) and some data items were not recovered. These items are not available for S34 data records where FDATE='31MAR2000' are: PRDATE COUNTRY TICKER2 EXCHCD STKCD SHROUT2 INDCODE SOVSHRS SHVSHRS and NOVSHRS.

Differences Between FDATE and RDATE, and the Late Report-Stale Data Issues

The two sets, S12 and S34, show holdings and related items for points in time that reflect both a 'vintage date' (FDATE) and a 'report date' (RDATE). The latter represents the date for which the holdings are valid (i.e., actually held by the manager or fund). FDATE signifies a particular vintage of data and serves as a primary key for joining certain tables that constitute the complete sets of S12 and S34 set.

FDATE and RDATE are the same date in a large majority of the Investment Companies in the S34 set. For mutual funds, a slight majority of the RDATEs are in the same quarter as the FDATE, and although reports may be made on any day, the last day of the quarter is most commonly the report day. Nonetheless, late reporters (i.e., cases where a fund rarely shows RDATE holding that correspond to the same quarter as the FDATE vintage) and stale data (i.e., where the same RDATE based holdings are shown in two or more consecutive FDATE vintages) are two important and problematic issues with the S12 set.

In order for the late reports to be identified with the right holdings date, we use RDATE to signify the date to which the items apply, after lining up the separate data tables using FDATE. Large gaps between FDATE and RDATE for any record (i.e., a fund or manager's holding) often corresponds to cases with large gaps in the sequence of RDATEs for a fund (FUNDNO) or manager (MGRNO), and thus missing holdings data. See the final section on 'Issues, Potential Problems, and Caveats' for details.

III. Mutual Fund Common Stock Holdings Database (S12)

A. Data source

The primary source for the mutual fund holdings data is SEC N-30D filings. These filings, which include semi-annual reports to shareholders, are required to be filed with the SEC twice a year by mutual fund companies.[4] To a lesser extent, Thomson taps fund prospectus and contacts mutual fund management companies to increase update frequency.

B. Coverage of Mutual Funds

The S12 mutual fund holdings database covers almost all historical domestic mutual funds plus about 3,000 global funds that hold a fraction of assets in stocks traded in U.S. exchanges as well Canadian stock markets. Because it keeps virtually all U.S.-based mutual funds in existence since 1980, this set is largely free of the survivor-bias that has been a major concern in the mutual fund research.[5]

C. Data Files

i. Source Files

The S12 source files are divided into distinct table 'types' that cover different aspects of mutual fund and their holdings. The items in each of these tables are described below, along with the role of a variable called FDATE that represents the file date, which WRDS adds to identify the 'cut' or vintage date.[6]

S12TYPE1—Fund Characteristics

The S12TYPE1 data reports mutual fund characteristics, including fund name (FUNDNAME), the corresponding fund number (FUNDNO), investment objective (IOC), total net assets (ASSETS), management company name abbreviation (MGRCOAB), country (COUNTRY), and the report date for holdings (RDATE).

The primary key for the TYPE1 file is FUNDNO, which serves as a numeric

[4] Prior to 1985, individual mutual funds are required to report portfolio holdings on a quarterly basis.

[5] See Appendix A of Wermers (1999) for more detail on this issue, the collection procedure adopted in creating the Thomson set and the structure of each table.

[6] For example, the table file named 'S121.0012' is a TYPE1 file for the S12 set. The file extension identifies the year as 2000 (yy=00) and the month as December (mm=12) for the file vintage, such that with FDATE='31DEC2000' in this case. These files were not actually cut or produced on the FDATE, but instead FDATE represents the quarter (last day of quarter) for which the data items were generally available for public information such as stock prices, and for holdings, theoretically available through fund or investment company records. In the WRDS system, the data in the S12n.yymm tables have been appended sequentially (i.e., in FDATE order) to create S12TYPEn tables.

identification variable for funds. In most cases, FUNDNO is a unique ID, but there are cases where a fund changes its name and Thomson assigns a new FUNDNO to the new name even though the fund's manager and shareholders remain unchanged. In addition, there are cases where a FUNDNO is reused, such that it is used to identify a new fund that is not related to the original and dead FUNDNO fund.

RDATE, which is the date for which holdings correspond, is often the same as FDATE, but can be any day in a quarter. In fact, this date can be in a prior quarter for reasons discussed below. Neither FDATE nor RDATE is the date that the report is filed with the SEC, which is not available in the any S12 (or S34) set.

Another important issue with the S12TYPE1 table is the fact that when a fund's holdings are not reported in the next quarter or if the holdings data for a fund fail to be collected for any reason, both RDATE and holdings data in other tables are carried forward to the next FDATE set, creating cases of 'stale' data. Therefore, RDATE in the TYPE1 table is used to make certain table joins, i.e. those that depend on the holdings date, while FDATE is used for other types of table joins, i.e. those that depend on the vintage date. (See below for a section that discusses the exact nature of these joins and the manner in which holdings values are reported in another table.)

S12TYPE2—Stock Characteristics

U.S. and Canadian stocks that are held by mutual funds are reported in the TYPE2 tables. For each stock that is held by at least one mutual fund in a quarter (identified by FDATE), Thomson reports its 8-character (alphanumeric) CUSIP, exchange ticker symbol (TICKER), an additional special ticker symbol (TICKER2),[7] company name (STKNAME), industry (INDCODE), stock code (STKCD), stock class description (STKCDESC), exchange (EXCHCD), quarter-end price (PRC), and quarter-end shares outstanding in both millions and thousands (SHROUT1, SHROUT2).

CUSIP is the primary key in these tables and FDATE identifies the date for which prices and shares outstanding are valid. As shown in the variables description table above, some of the variables were not part of the original CDA/Spectrum design and become available in either 1994 or 1999.

PRC, SHROUT1 and SHROUT2 are usually adjusted for stock splits and other corporate events in the quarter, in order to represent actual trading price and shares outstanding at the end of the quarter (FDATE). Because these adjustments cannot be synchronized with every RDATE in the S12TYPE1 files, they are not proper for all cases. (This issue is discussed in detail in the TYPE3 description.) There are also cases where an adjustment is made for a stock split that actually occurs after the FDATE. This asynchrony in adjustments also occurs to shares held by funds in the TYPE3 file and change in number

[7] TICKER starts to be available from March 1980 with a length of four characters and TICKER2 from March 1994 with a length of five characters.

of shares in the TYPE4 file. In almost all quarters, if shares outstanding are adjusted for a stock-quarter, the shares and change in shares of that stock held by mutual funds in that quarter are adjusted by the same proportion.

S12TYPE3—Holdings

The TYPE3 file provides the number of SHARES (by CUSIP) that each mutual fund (FUNDNO) held, starting with the first quarter of 1980. The date for which the holdings are valid (RDATE) is not identified in this table, but instead must be found by linking to FUNDNO-FDATE in the TYPE1 table and reading the corresponding RDATE.

Holding adjustments are made for stock splits, stock distributions, mergers and acquisitions and other corporate events, such that the SHARES values are adjusted for stock splits that occur between the linked RDATE and FDATE. Presumably, the mutual fund's reported holdings would not include such adjustments. This adjustment is needed to properly compute the value of the holdings and the percent of total shares outstanding from the PRC, SHROUT1 and SHTOUT2 values in the S12TYPE2 tables. For example, holdings data for a fund with a RDATE=30SEP1999 in the FDATE=31DEC1999 that shows holdings on September 30, 1999 have been adjusted for any stock splits between this date and December 31, 1999. There may be problems with split adjustments for such late reporters, however, because of the manner in which Thomson applies splits, or specifically that the assumption by Thomson that the reported shares held are pre-adjustment may not be correct in all cases. (We are looking into ways to fix these problems).

S12TYPE4—Change in Holdings

The TYPE4 file reports the CHANGE in the number of shares held and the associated key variables are FUNDNO and CUSIP. To calculate a mutual fund's net transaction on a stock, Thomson subtracts the number of shares held at the *current report date* (RDATE) by the number of shares held at the *prior report date* (PRDATE), although neither of these variables are in the TYPE4 set (they are inferred by linking to other sets using linking keys as described below). The CHANGE is in split-adjusted terms, to be consistent with the SHARES value in the S12TYPE3 set. In other words, CHANGE will not equal the reported difference in SHARES from two consecutive reports if the prior report values are in pre-adjustment terms, but will equal the difference in SHARES from two consecutive reports if the prior report values is adjusted to reflect any split effects that are incorporated into the more recent report (as identified by the FDATE-RDATE combination associated with SHARES).

S12TYPE5—Mapping with S34 Institutional Money Managers (Starting from 1994)

In the TYPE5 table, fund numbers (FUNDNO's) found in the mutual fund holdings database are mapped to the manager numbers (MGRNO's) found in the institutional

holdings database on a quarterly basis. Thomson began to report these FUNDNO-MGRNO links in the first quarter of 1994 but for numerous applications, the first (1994:Q1) link for a FUNDNO can be used with the 1980-1993 observations.

S12TYPE6—Permkeys (Starting from 1994)

Table S12TYPE6 maps the older legacy permanent key numbers (PERMKEY) to the newer fund numbers (FUNDNO). Permanent keys were used by Thomson at one time instead of fund numbers. The S12TYPE6 table reports observations starting from the first quarter of 1994 but was not consistently updated, especially after 1999, and therefore has little value.

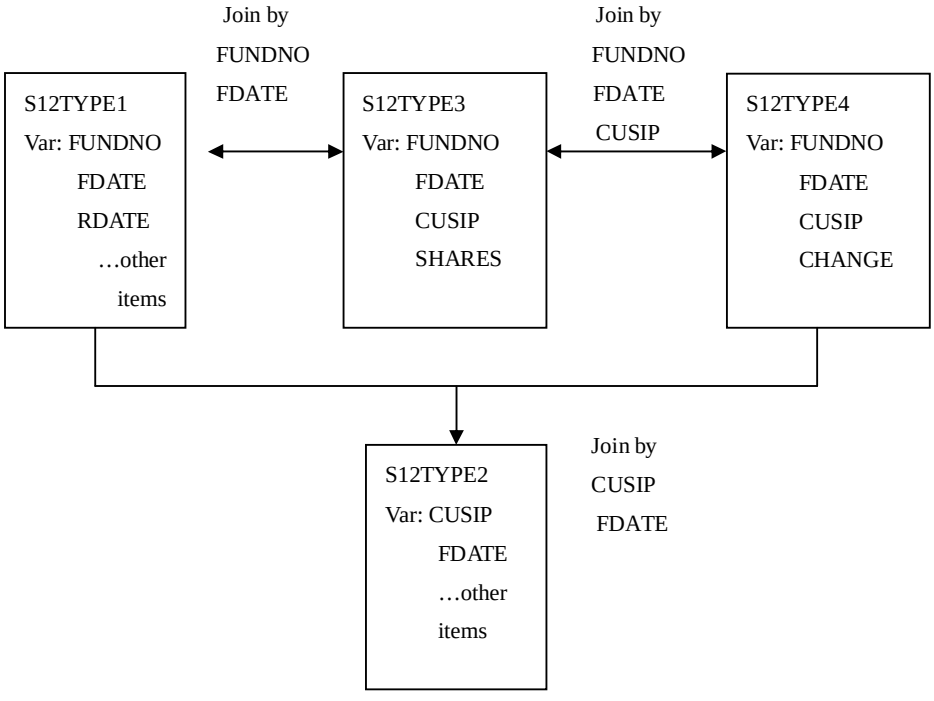
S12TYPE7—Management Company Names (Starting from 1999) Table S12TYPE7 provide a mapping from "Management Company Name Abbreviation" (MGRCOAB) reported in the TYPE1 file to the full "Management Company Name." The S12TYPE7 table reports observations starting from the first quarter of 1999.

S12TYPE8—Fund Ticker Symbols (Starting from 1999)

Fund ticker symbols are reported in the S12TYPE8 table on a quarterly basis starting from the first quarter of 1999. For fund portfolios offering multiple share classes, multiple ticker symbols are provided.

ii. **S12 – Join of TYPE1, TYPE2, TYPE3, and TYPE4 items**

Although the design of the 'type' files is sensible as a basic database structure for archival purposes, this set up is not handy for searching and extracting holdings records for a particular fund during a particular quarter because the necessary information is scattered across the four tables or files. generates a comprehensive SAS data set named S12 using the following scheme.



As shown in the above chart, the four tables are joined by a common key variable and a date variable -- the FDATE.

iii. **S12NAMES**

The ‘names’ file for the S12 set, *S12NAMES*, contains a compete list of fund identifiers (FUNDNO) and names (FUNDNAME), along with management company abbreviations (MGRCOAB), country (COUNTRY) and investment objective codes (IOC). This file is basically a summary of the S12TYPE1 set, such that only one record for each FUNDNOFUNDNAME combination is shown. It is most useful for determining the FUNDNO identifier for a particular fund name, or alternatively, determining the historical names for a FUNDNO. The table is sorted by FUNDNO and RDATE1, which denotes the first report date for the FUNDNO-FUNDNAME pair. Also included is RDATE2, the last report date for the FUNDNOFUNDNAME pair. Note that more than one row for a FUNDNO will be shown if there was a name change, which may be a case of a different abbreviation in the S12 source data, as opposed to an intentional name change by the fund manager.

IV. Investment Company Common Stock Holdings Database (34)

A. Data source

The primary source for the institutional holdings data is the 13f form that investment companies and professional money managers are required to file with the SEC on a quarterly basis.

B. Coverage of Institutional Money Managers

Under Securities Exchange Act Section 3(a)(9) and Section 13(f)(5)(A), an institutional investment manager is an entity that either invests in, or buys and sells, securities for its own account; or a natural person or an entity that exercises investment discretion over the account of any other natural person or entity. Currently, only managers with over $100 million under their control are required to file, but others they may still do so and be included in the S34 set.

The S34 holdings (identified by CUSIPs) are generally equities, but are not necessarily the entire equity holdings of the manager. For one, small holdings—under 10,000 shares or $200,000—may be omitted from 13f reports, as well as cases where there may be confidentiality issues. In addition, the CDA and TFN criteria omit some equity holdings that could not be matched to a master security file. This problem is likely to be small in recent years, and more likely for foreign equity holdings. Fixed income holdings are not intentionally included, but the holdings may include cash- like holdings in cases where the reported CUSIP could be matched to a security master file. There are also cases where two or more managers have shared, discretionary control over a holding, but potential duplicate counting is mitigated by an SEC rule that requires only one manager to include the holdings information in their 13f report.

A relationship between the S12 set and the S34 set comes from the fact that each fund has a manager, and in most cases the managers are in S34 set. For example, Fidelity (MGRNO=27800 in the S34 set) reports as a single managing entity, aggregating the holdings of all funds in the Fidelity family (such as Magellan, FUNDNO=21858 in the S12 set). Fidelity consistently reports it manager holdings on a quarterly basis, however, compared to the individual mutual fund holdings that are only required to be reported semi-annual. In addition, a mutual fund manager's holdings in the S34 set may be greater than a simple aggregate of its funds because it manages investment vehicles for trusts, pensions, and individuals that are technically not mutual funds.

C. Data Files

i. Source Files

As with the S12 set, the S34 source files are divided into distinct table 'types' that cover

different aspects of mutual fund and their holdings. The items in each of these tables are described below, along with the role of a variable called FDATE that represents the file date, which WRDS adds to identify the 'cut' or vintage date.[8]

S34TYPE1—Institutional Manager Characteristics

The TYPE1 table of the S34 set provides information on institutional money managers' characteristics such as name (MGRNAME), COUNTRY, a legacy identifier (PERMKEY) and manger type (TYPECODE).[9] Three date variables, file vintage (FDATE), holdings report date (RDATE), and the prior holdings date report (PRDATE) are also included in this table. This table is sorted by MGRNO and FDATE.

S34TYPE2—Stock Characteristics

Information about stocks held by the institutional money managers is contained in the S34TYPE2 table. This table is basically the same as the S12TYPE2 table described above. In terms of stock coverage, S34TYPE2 has a larger number of stocks, as mutual fund stock holdings are a subset of the stock holdings by all types of institutional money managers.

S34TYPE3—Holdings

S34TYPE3 contains holdings information such as stock CUSIP (CUSIP), manager number (MGRNO), type code (TYPE), and shares held at quarter end (SHARES). One difference between S34TYPE3 and S12TYPE3 is that from 1999, Thomson offers a breakdown of the number of shares held based on whether and how much an institution has investment discretion/voting authority. Specifically, three variables—NO, SOLE, and SHARED—are used to record the number of shares held for which the filer exercises no voting authority, sole voting authority, and shared voting authority, respectively. By definition, the sum of the three variables should be equal to SHARES.

S34TYPE4—Change in Holdings

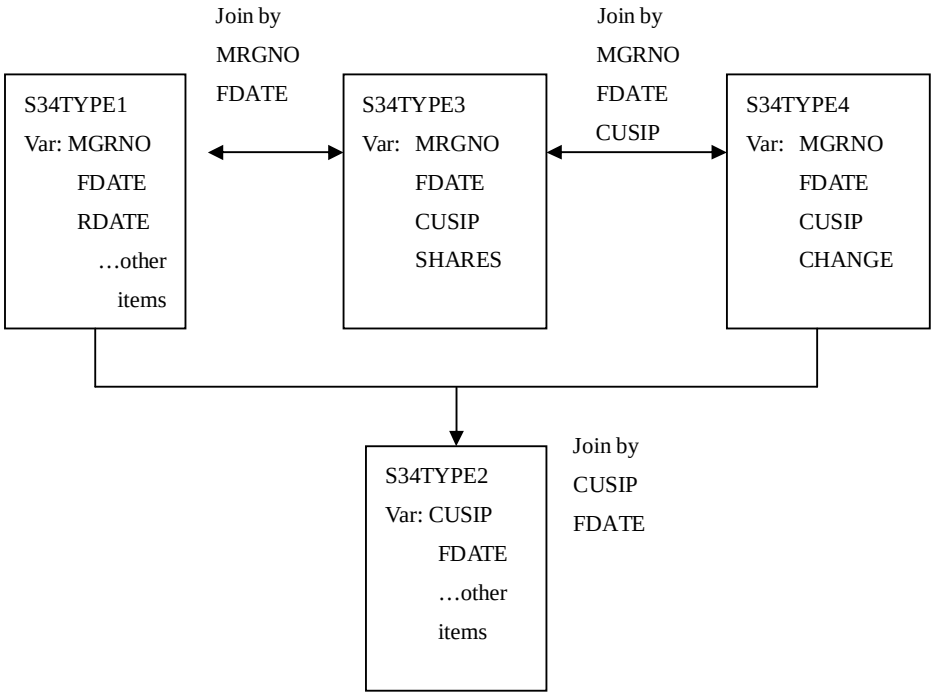
Stock transactions of institutional money managers are recorded in the TYPE4 file, which contains the following four variables: 8-digit stock CUSIP (CUSIP), manager number (MGRNO), manager type (TYPE), and net changes in shares since prior report (CHANGE).

[8] FDATE has the exact same properties as described in the S12 documentation above.

[9] The TYPECODE variable, renamed as TYPE1 in the all-inclusive S34 set, was designed to distinguish among different types of institutional managers. It has the problem of sudden change from a non-five value to five in December 1998, March 1999, and June 1999.

ii. ***S34***

WRDS merges the four S34 TYPE files to generate an all- inclusive, comprehensive SAS data set named S34 using the following scheme.



The join of the four tables (as shown in the chart) is completely analogous to the method described above for the S12. The only difference is that MGRNO replaces FUNDNO as the entity identifying key for joining the TYPE1, TYPE3, and TYPE4 tables.

ii. ***S34NAMES***

The 'names' file for the S34 set, *S34NAMES*, contains a compete list of management company identifiers (MGRNO) and names (MGRNAME), along with management type code (TYPECODE) and COUNTRY (country). This file is basically a summary of the S34TYPE1 set, such that only one record for each MGRNO-MGRNAME combination is shown. It is most useful for determining the MGRNO identifier for a particular investment company name, or alternatively, determining the historical names for a MGRNO. The table is sorted by MGRNO and RDATE1, which denotes the first report date for the MGRNO-MGRNAME pair. Also included is RDATE2, the last report date for the MGRNO-MGRNAME pair. Note that more than one row for a MGRNO will be shown if there was a name change, which may be a case of a different abbreviation in the S34 source data, as opposed to an actual change in the investment company's name.

V. Issues, Potential Problems for Users, and Caveats

1. Coverage and Missing Holdings Data

The holdings in the S12 and S34 sets are rarely the entire equity holdings of the manager or fund. There are minimum size requirements and confidentiality qualifications. In addition, some equity holdings were dropped if Thomson could not match the reported CUSIPs to their master file of securities, which is more likely the case for foreign equity holdings. This problem with unmatched and dropped holdings is likely to be small in recent years, however.

A bigger problem comes from late reporting and 'stale' data, which can lead to gaps in the holdings data. As explained above, the holdings and related items are 'marked' for points in time that reflect both a 'vintage ' date (FDATE) and a 'report' date (RDATE), and the latter date is the appropriate one to use for designating the date of the holdings.

Late reports and stale data are reflected in cases where RDATE and FDATE are no t in the same quarter and the gap between these two dates is irregular. While it is somewhat common in the S12 source files to find the exact same holdings data for two consecutive vintages, both representing the same RDATE, simply because mutual funds are only required to report semiannually, there are also cases where the FDATE-RDATE patterns are clearly not consistent with either quarterly or semi-annual, or even an alternating combination of the these two types of reporting frequencies reports.

For example, there are cases where the S12 source data show a new set of holdings on quarterly basis for a majority of a fund's history and then has an unexplained gap, where a sequence of three or more FDATE vintages shows the same RDATE and holdings data. Below is an example that shows inconsistent numbers of months between reports and one large gap between reports.

FUNDNO=1074, JOHN HANCOCK FINANCIAL INDUSTRIES FUND

FDATE	RDATE		Months Since Prior Report
30Sep1996	31Jul1996	*	—
31Dec1996	31Oct1996	*	3
31Mar1997	31Mar1997	*	5
30Jun1997	30Apr1997	*	1
30Sep1997	30Apr1997		—
31Dec1997	31Oct1997	*	6
31Mar1998	31Oct1997		—
30Jun1998	30Apr1998	*	6
30Sep1998	30Sep1998	*	5
31Dec1998	31Dec1998	*	3
31Mar1999	31Mar1999	*	3

30Jun1999	31Mar1999		—
30Sep1999	30Apr1999		1
31Dec1999	30Apr1999		—
31Mar2000	30Apr1999		—
30Jun2000	30Apr1999		—
30Sep2000	30Apr1999		—
31Dec2000	31Dec2000	*	20
31Mar2001	31Mar2001	*	3
30Jun2001	30Jun2001	*	3
30Sep2001	30Sep2001	*	3
31Dec2001	31Dec2001	*	3
31Mar2002	31Dec2001		—
30Jun2002	30Jun2002	*	6
30Sep2002	30Sep2002	*	3
31Dec2002	31Dec2002	*	3
31Mar2003	31Mar2003	*	3
30Jun2003	30Jun2003	*	3
30Sep2003	30Sep2003	*	3

* Case where RDATE and FDATE are in the same quarter.

In the table above, the underlined dates denote cases where the FDATE denoted vintage is the first report for the corresponding RDATE. For this fund in the 1996 to 1998 period, the gap between reports was not consistent: three times it was 3 months, two times it was 6 months, two times it was 5 months, and in one case it was only 1 month. There are also two vintage reports in this period that we ignore in constructing the S12 set (FDATE = 30Sep1997 and FDATE =31Mar1998) because the same RDATE designate holdings were already recorded in the prior vintages. The bigger problem is that there were only two reports in 1999 and there were only one month apart. After the April 1999 report, the next effective report is for December 2000, creating a gap of 20 months. From 2001 forward, however, the reports are regularly spaced at 3 months, save for one case where there is no March 2002 report.

The main problem is that due to the manner that the CDA and TFN sets were constructed and archived there is no mechanism for allowing late reporters to ‘catch up’ once a report is missed or late. If a manager or fund subsequently makes two reports (late and most recent) in the same quarter, it is likely that there will be a gap in the holdings data.[10]

2. Other Known Anomalies

[10] The CDA/Thomson documentation for the construction of the source data is sketchy, but it likely that only the last report in a quarter was kept. It is also possible that the fund or manager made all reports on time, but CDA or Thomson did not properly update the source files. In both cases, the main issue or mistake was that the records in the prior (FDATE vintage) files that carried stale holdings were not replaced with updated data and the consequences can not solvable by changing the table joining scheme.

There are various anomalies and data problems in the TFN S12 and S34 sets. Some of the problems are being resolved, but others cannot be fixed because the historical data was lost or corrupted. Thomson explains that "[t]he historical archive files have been produced over a long time frame and have been modified over the years. Several data items have been added as they became available, such as the historical files differ since the more recent files have additional data included."

This means that various data items show inconsistent coverage and format over time. Thus, all users need to verify that it meets their needs and we believe they will need to perform some data cleansing that is specific to their application.

An important fact to know is that the FUNDNO and MGRNO numbered identifiers are reused in the TFN sets and do not serve as unique and permanent identifiers for every fund or manager. A gap of more than one year in the RDATE for the same FUNDNO or MGRNO typically reflects a different and unrelated fund or manager. In other cases, a fund or manager may be reassigned a different FUNDNO, usually when a name changed occurred. Most FUNDNAME and MGRNAME changes are not associated with a different number, however, but instead reflect an official name change or a different manner of abbreviating the same name (i.e ALLEN B. SMITH INVESTMENT MANAGEMENT -> AB SMITH INVT MGMT CO).

Therefore a terminal RDATE for a fund or manager does not necessarily signify that the fund or manager is no longer operating. Similarly, following a particular FUNDNO or MGRNO over time may not provide data on the same manger or fund. It is recommended that Manager Names and Fund Names be studied before drawing conclusion about particular cases. PERMKEYs are not viable alternative identifiers for most cases, since they were first used in 1994 and are not consistently populated in recent years.

TYPECODE in the S34 set have serious classification errors in recent years, such that the by the Other group is unrealistically large. Many Banks (TYPECODE=1) and Independent Investment Advisors (TYPECODE=4) are improperly classified in the Others (TYPE=5) group in 1998 and beyond. For example, in the first quarter of 1999, the number of independent investment advisors drops from over 1200 to about 200, while the Other group jumps from roughly 100 to over 1300 . TFN explain that a mapping error occurred when integrating data from another source, regret that the problem occurred, but they have no plans to fix the problem.

A problem in the original data files with CUSIPs having no leading zeros in records prior to 1988 has been fixed in SAS datasets on WRDS.

EXCHCD and STKCD were placed in 1993:4 (for S12 set) and 1994:1 (for S34 set) files but are not are consistently reported until 1999. STKDESC is only reported for the S12 set in 1999.

Other problems are most prevalent in the 1999 and 2000 data files, and may be related to the fact that Thomson modified the format and items included in the basic data files in 1999. For example, December 1999 fund coverage ended with FUNDNO=77700 in the original data files provided to WRDS. A subsequent fix was made, but there is no guarantee that funds numbers higher than 77700 are not missing at least one report for 1999 or 2000.

Data coverage in March 2000 and June 2000 was originally different from the rest of the data due to lost files. For the most part, the problem was fixed with replacement files from Thomson for the S12 set, but missing items remain for the S34 tables where FDATE=31MAR2000: PRDATE (Prior Report Date), COUNTRY (Country), TICKER2 (Extended Company Ticker Symbol), EXCHCD (Exchange Code), STKCD (Stock Class Code), SHROUT2 (End of Qtr Shares Outstanding in 1000s), INDCODE (Industry Code), SOVSHRS (Shares with Sole Voting Authority), SHRSHRS (Shares with Shared Voting Authority), and NOVSHRS (Shares with No Voting Authority).

Prices and Shares Outstanding in the 1999 and 2000 files show cases where updates were not made—i.e. same value as the prior quarter. The problem is most prevalent with the December 1999 stock prices for both the S12 and S34 sets. A similar problem is seen with the September 2000 prices for the S34 set.

References

Gompers, Paul, and Andrew Metrick, 2001, "Institutional Investors and Equity Shares," The Quarterly Journal of Economics, p.229-259.

Wermers, Russ, 1999, "Mutual Fund Herding and the Impact on Stock Prices," Journal of Finance 54(2), p.581-622.

Wermers, Russ, 2000, "Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transaction Costs, and Expenses," Journal of Finance 55(4), p.1655-1695.

This document was written by Michael Boldin (WRDS) and Bill Ding (visiting WRDS in 2003, currently at SUNY— Albany), and it replaces a similar document dated August 2003. It is based upon various Thomson documents, with additional information and documents provided by Bharat Daphtary, David Granger, Ben Norton, and Nathan Wolaver at Thomson. Russ Wermers (University of Maryland) and Andrew Metrick (Wharton) were also helpful in providing information about the structure of the original CDA data and source files.

Appendices

1. Code Definitions

a. Investment objective code (variable name: IOC; SAS dataset: S12TYPE1)

Code	Investment Objective
1	International
2	Aggressive Growth
3	Growth
4	Growth & Income
5	Municipal Bonds
6	Bond & Preferred
7	Balanced
8	Metals
9	Unclassified

b. Type Code (Variable Name: TYPECODE; SAS dataset: S34TYPE1)

Code	Institutional Type
1	Bank
2	Insurance Company
3	Investment Companies and Their Managers
4	Investment Advisors
5	All Others (Pension Funds, University Endowments, Foundations)

c. Exchange (Variable Name: EXCHCD; SAS datasets: S12TYPE2, S34TYPE2)

Code	Exchange Name
A	NYSE
B	AMEX
C	PBW & Boston
D	Midwest
E	Pacific
F	NASDAQ Small
I	OTC Mutual Funds
J	Toronto
K	Montreal
V	NASDAQ National
W	Vancouver
Z	Alberta

d. Stock Class (Variable Name: STKCD; SAS datasets: S12TYPE2, S34TYPE2)

Code	Stock Class
0	Common stock
4	
7	Mutual Fund
8	
+	Foreign stock
-	

e. Stock Description (Variable Name: STKCDESC; SAS datasets: S12TYPE2, S34TYPE2)

Code	Stock Description
AR	
CMA	Common Class A
CMB	Common Class B
CMC	Common Class C
COM	Common Stock
LTD	
OR	Ordinary Share
PR	Preferred Stock
RG	Right
RT	Receipt
UNT	Unit

f. Industry code (variable name: INDCODE; SAS datasets: S12TYPE2, S34TYPE2)

Code	Industry
100	Unknown
101	Aerospace
102	Agriculture
103	Airlines
104	Automobiles
105	Banks & Savings Institutions
106	Beverages
107	Chemicals
108	Computer Hardware Software & Services
109	Construction & Engineering
110	Consumer Services
111	Electrical & Electronics
112	Miscellaneous
113	Energy and Fuels
114	Financial Services
115	Food & Restaurants
116	Healthcare

117	House Wares & Household Items
118	Industrial Manufacturing
119	Insurance
120	Investment Services
121	Leisure Travel & Lodging
122	Machinery & Equipment
123	Media
124	Metals & Mining
125	Packaging
126	Paper & Forest Products
127	Publishing & Printing
128	Real Estate
129	Retail & Consumer Goods
130	Semiconductors
131	Telecommunications
132	Textiles & Apparel
133	Tobacco
134	Transportation
135	Utilities: Water-Electric-Gas
136	Waste & Environment Management