User's Guide

WizSame®

Version 2014

WizSoft®

1. Welcome!

WizSame is a tool for revealing duplicate records, or similar records suspected as being duplicate.

Who are *WizSame* users?

Revealing duplicate records is relevant to two kinds of users:

(1) Auditors search for duplicate records in order to reveal errors and frauds. Usually they search for duplicate records such as:

- Duplicate entries for the same customer / employee / stock item

- Duplicate invoices for the same purchase as well as duplicate payments

Obviously such duplicate records are cases to be audited.

(2) Marketing and CRM managers search for duplicate records when merging a new list of potential customers in the existing customer database. They are interested in revealing existing customers in the new list and deleting them from merged new customers.

WizSame meets both of these requirements.

Similar records suspected as duplicate

Obviously identical records are duplicates, but in many cases the duplicate records are not identical but just similar. *WizSame* reveals similar records suspected as being duplicates.

There are several ways in which one record may be similar to another:

(1) The two records may be similar since they contain similar names, for example, in one record the customer name is *Gurbatchev* while in the other record it is *Gorbatchev*. *WizSame* reveals all the cases where two names differ by one character. Such a difference can be materialized in three ways:

    a. One character in one record is replaced by another character in the second record, in the same place. For example, the second character, *u*, in *Gurbatchev* replaces the second character, *o*, in *Gorbatchev*.

    b. One character is included in one name and is absent in the other name. For example, *Gurbatchev* and *Gaurbatchev* – the second character, a, is included in the first name only.

    c. Two adjacent characters in one name appear in an opposite order in the other name. For example, *Gurbatchev* and *Grubatchev* – the second and the third character appear in opposite order in these two names.

(2) Several records may be similar when they contain synonymous names, for example in one record the state is *N.Y.* while in the second record the state is *NY*, and in the third record the state is *New-York*. *WizSame* contains a user-updateable synonym dictionary where each user can enter as many synonyms as he or she needs (in addition to the synonyms that are already included in the *WizSame* dictionary).

(3) Records may also be similar when they contain the same values in another order (in the same field). For example, in one record the customer name is *George Ernst*, while in the other record it is *Ernst George* – same names, different order. *WizSame* reviews all the possible comparisons between the strings in each field.

(4) Finally several records may be similar if they have identical or similar values in one field *or* in another. For example, one may define records as similar if either the customer email address *or* the phone number is identical. *WizSame* lets the user define the several conditions connected by the AND or OR operators to determine matching.

How does *WizSame* deal with numbers?

When searching for similarity *WizSame* reads numbers as if they are names. That is, two numbers are considered similar when they differ by one character according to the above-mentioned criteria. For example the following numbers are similar to the first one:

123.45

123.46 (replacement of one digit)

7,123.45 (one extra digit)

123.54 (same digits, different order of the last two)

What are the parameters for defining similarity?

The user determines for each field whether the field should be –

- Identical

- Similar

- Ignored

The user may also define conditions, connected by the AND or OR operators, such as: (Several records are identical if:) the values in Field A are similar OR the values in Field B are identical.

What do I do with the discovered duplicate records?

As mentioned, *WizSame* addresses two kinds of users:

- Auditors looking for duplicate customers, invoices payments, etc.;

- Marketing and CRM managers looking for existing customers in a new list of potential customers that supposed to be embedded in the existing database.

When auditors use *WizSame* they run the program on the table under investigation. *WizSame* displays a report containing all the discovered matching sets. When reviewing this report the user can –

(1) Review the matching sets on the screen.

(2) Print a report of the matching sets.

(3) Create a text file containing all the matching sets including the data set number and the matching set number.

(4) Search for a certain record to see in what matching sets, if any, the record is included. This option is useful when one checks whether a certain record has duplicates.

(5) Select a group of matching sets to be printed as a text table or saved as a text file. This option is useful when the user wishes to reduce the number of matching sets by applying conditions such as: selects all the matching sets where Field A values starts with 212.

When marketing and CRM managers use *WizSame* they run the program on two tables – one table containing the existing database, and a second table containing the new list. *WizSame* reveals records in the new list that are identical or similar to records in the existing database. *WizSame* lets the user delete the duplicate records from the new list by creating a text file that contains only the non-duplicate records, which can then be merged with the existing data.

The contents of this guide

At the end of the manual:

The **Quick Reference** appendix describes the main *WizSame* functions organized according to the *WizSame* menu options. It should prove useful after you are "up and running," as a fast reminder of how to perform certain operations.

2. Installing *WizSame*

WizSame can be installed to operate as a stand-alone system or as a network application. A CD containing both the *WizSame* application and the software for installing ODBC and OLE DB drivers is included in the installation kit.

Before you begin installation, be sure to close all open applications.

Software Installation

Insert the CD into the appropriate drive of your computer. Usually the installation will start automatically. If not,

- From the Start menu, select Run to display the Run dialog box.
- In the Open text box, select d:\setup or e:\setup, as appropriate, and click OK.

The Master Setup Dialog window will be displayed, presenting the following installation options:

- Click on the Install WizSame button to install the *WizSame* program.
- The Install Microsoft Internet Explorer ver. 5 button is active if the Microsoft Internet Explore ver. 5 (or later) is not already installed. This program is needed for the *WizSame* online Help.

After installing one program you can repeat the above steps to install another.

Read the displayed Welcome screen and click the Next command button when you are ready to continue. The Registration window will be displayed.

Enter your Name and Company, click Next and then confirm your selections by clicking Yes. The Setup Install Type dialog box will be displayed.

Select the type of *WizSame* installation that you require. First-time users should usually select the Typical option.

In the Destination Directory block, check the displayed path for installing *WizSame*. If it is not appropriate, use the Browse button to indicate the desired path. Click Next to continue.

Follow the displayed instructions to complete the installation process.

3. Getting Started

Before you begin using *WizSame*, be sure that you have a data set containing the data you wish to analyze.

The *WizSame* Menu

File	Edit	View	Issue	Settings	Window	Help
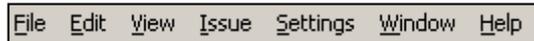
The *WizSame* menu contains the following options:

   File - for managing the *WizSame* files and for opening data tables.

   Edit - for performing common editing functions.

   View - for displaying or hiding the *WizSame* tool bars.

   Issue - for printing all the matching sets, filtering the matching sets, print to a new file

   Settings - for selecting fonts.

   Window - for selecting the active window.

   Help – for Help on using *WizSame*

The Action Bar

Issue Report

Click the Issue Report button to instruct *WizSame* to reveal the records suspected as duplicate records and issue the reports.

The Report bar

You can use these three command buttons to move between *WizSame*'s main window and the report windows.

Click on the left button to move to the main window.

Click on the button with the R to move to the report.

The Window bar

You can use this Window bar to move between *WizSame*'s main window and the report windows like the Report bar.

Click on the **+** and then select the window.

Sequence of operations

There are four main steps in operating *WizSame*:

First, you open the data set to be analyzed. You have several options for this step. *WizSame* can --

- Read a text file using the *WizSame* ASCII reader.
- Directly read a file with a .dbf extension.
- Directly read MS Access table(s)
- Directly read MS SQL table(s)
- Directly read an Oracle data set.
- Read a data set (having one or more tables) through ODBC.
- Read a data set (having one or more tables) through OLE DB.

Once the data set is open, you determine the criteria for selecting records into matching sets, where the records in each matching set are suspected

to be duplicates. For each field you determine whether the values of the duplicate records should be identical or similar.

Click on the Issue Report button. *WizSame* will display the matching set report.

You can then filter the matching sets and issue a new file containing no matching sets.

4. Tutorial

This chapter teaches you how to perform the basic *WizSame* operations. In this tutorial, you will:

- Open a data set.

- Define the analysis parameters.

- Review the reports.

Before you begin, be sure that your system has been properly installed and that WizSame example.mdb resides in your WizSame folder. The table contains sale transaction data. You will apply *WizSame* on this table in order to reveal duplicate records.

Access the *WizSame* application by double clicking on the *WizSame* icon on the desktop. Alternatively, open the Start menu and from the Programs list, select WizSoft Applications and then select WizSame. The *WizSame* desktop will be displayed.

Opening the data set

Click on the Open Data of Type button. The Open dialog box will be displayed.

Click on the WizSame Example data source name and click OK. The MS Access Data Source dialog box will be displayed.

Select the WizSame Example data source in the right pane. Click OK. *WizSame* will open the table and display its fields in the *WizSame* work area.

The directory and name of the table that you selected is displayed in the Data Source text box. Under that is the field grid, which includes the names of all the fields listed in the Field Name column.
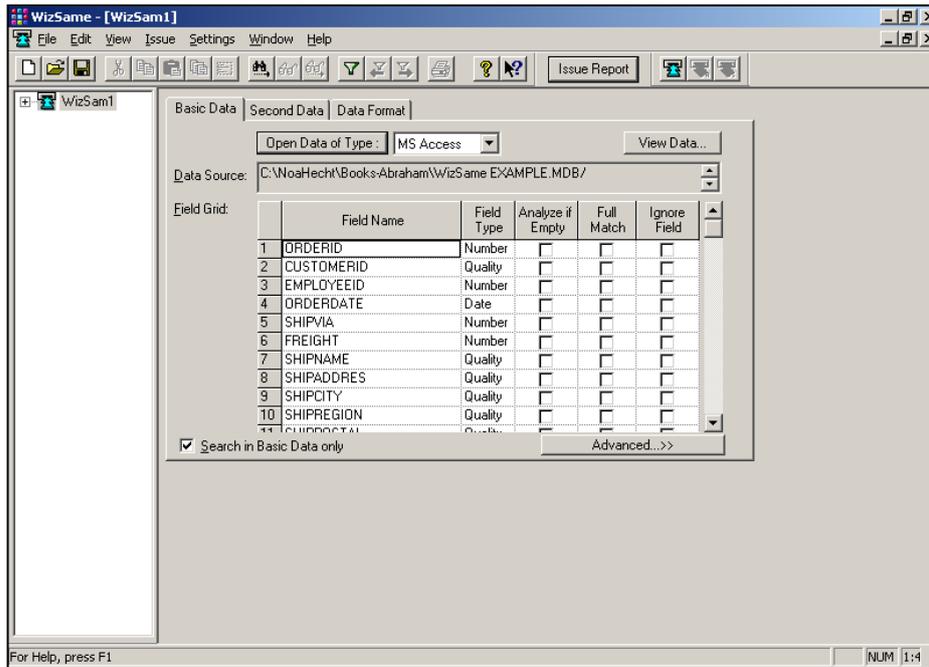
Determining the analysis parameters

Once the data set to be analyzed is open, the analysis parameters can be determined. However in this tutorial you will leave the default parameters "as is". (You can read about these parameters in chapter 6).

Click on the Issue Report button.

Reviewing the reports

You can now review the report (read chapter 7 for more details).

The screen displays two similar records suspected as duplicate. A checkmark denotes identical values, while a checkmark with an x signifies similar values.

5. Opening the Data Set

WizSame can read seven types of data

- Files produced through dBase, FoxPro, Clipper etc. (*.dbf files) are read directly.

- Microsoft Access (*.mdb) tables are read directly.

- Microsoft SQL Server tables are read directly.

- Oracle data tables are read directly.

- ODBC (Open DataBase Connectivity) compliant databases (produced by Access, SQL, Oracle, Sybase, Informix, DB/2 and other programs) can be read through the ODBC interface.

- OLE DB compliant databases (produced by Access, SQL, Oracle, Sybase, Informix, DB/2 and other programs) can be read through the OLE DB interface.

- ASCII-type text files can be read through the *WizSame* ASCII reader.
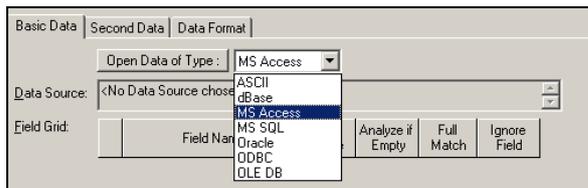
WizSame can read Microsoft Access, Microsoft SQL and Oracle data both directly, and through its ODBC or OLE DB utility. The direct reading is simpler.

Direct reading of *.dbf files

To open a *.dbf file -

- In the Open Data of Type list box select dBase, or click on the Open Data of Type button if dBase has been already selected. The Open dialog box will be displayed.

- In the Files of type text box at the bottom, be sure that the option dBase, FoxPro, Clipper Files (*.dbf) is displayed.

- Using the Look in list at the top, locate the directory containing the database file. All *.dbf files in that directory will be displayed.

- Select the file to be analyzed from the list.

- Click Open. The data set you selected will be displayed in the *WizSame* work area.

Direct reading of a single Microsoft Access (*.mdb) table

WizSame can read one or several tables. In the following lines, the instructions for how to open a single table are presented. The instructions how to join several tables can be found in pages 20 - 22.

- To open a single Access table, in the Open Data of Type list box select Access, or click on the Open Data of Type button if Access has been already selected. The MS Access Open dialog box will be displayed.

- Select ACCESS Files (*.mdb) from the Files of type drop-down list. Click on the file name and then click Open. The Access Data Sources dialog box will be displayed.

- Click on the data source name. The tables of this data source will be displayed in the right pane.

- Select the table to be analyzed in the right pane and click OK.

Direct reading of a single Microsoft SQL table

WizSame can read one or several tables. The instructions for how to open a single table are presented below. The instructions how to join several tables can be found in pages 20 - 22.

- To open a single Microsoft SQL table, in the Open Data of Type list box select MS SQL, or click on the Open Data of Type button if MS SQL has been already selected. The Microsoft SQL Login box will be displayed.

- Select or type the server service name in the Server box.

- Enter the User Name and Password for this service, or select the Trusted Connection check box if you are a member of the SQL Server users (your Windows Login Username & Password will be used).

- Click the Connect button. The MS SQL Data Sources dialog box will be displayed.

- Locate your database tables. The database tables will be displayed in the Contents list in the right pane of the SQL Data Sources dialog box.

- Select the data set table(s) to be analyzed in the right pane and click OK.

Direct Reading of a single Oracle table

WizSame can read one or several tables. The instructions for how to open a single table are presented below. The instructions how to join several tables can be found in pages 20 - 22.

- To open a single Oracle table, in the Open Data of Type list box select Oracle, or click on the Open Data of Type button if Oracle has been already selected. The Oracle Login box will be displayed.

- Select or type the server service name in the Server list box.

- Enter the User Name and Password for this service.

- Click the Connect button. The Oracle Data Sources dialog box will be displayed.

- Locate your database tables. The database tables will be displayed in the Contents list in the right pane of the Oracle Data Sources dialog box.

- Select the data set table(s) to be analyzed in the right pane and click OK.

Combining several tables into one *WizSame* data set

You can combine two or more tables by defining the relationships between them. The Database Relationships window is displayed automatically when multiple tables are selected. Two means of definition are available:

- Defining the table relationships directly (text view)

- Defining the links graphically

Defining relationships directly

- Select the Text View tab.

- In the displayed dialog box, select the Primary Table and Primary Field from the first two lists. The primary table and field(s) are the basis from which the relationships are defined. These connections can be either from the primary table/field to a single related table/field (one-to-one) or from the primary table/field to a number of other related tables/fields (one-to-many).

- From the subsequent lists, select the Related Table and Related Field and click Add Join. The related fields will be displayed in the list at the center of the dialog box.

- Repeat the previous steps to define all the relationships. You can use the command buttons at the right as follows:

To *change* a defined relationship, select the defined relationship in the list, revise the relationship by redefining the Primary Table/Field and the Related Table/Field and then click Change.

To delete a relationship, select it from the list and click Delete.

- With Access files, you may have *WizSame* define relationships automatically by clicking Autodetect. The system will build the relationships according to identical field names in the tables *and* field types. If a relationship between two tables of non-similar names and types has been defined previously, Autodetect will not delete the relationship and will not create any new relationship between those tables.

- When reading tables directly (*not* through ODBC), *WizSame* will identify relationships defined previously in the database itself.

- To revise your selection of tables, click the Tables button. The Tables dialog box will be displayed.

- The Selected Tables list at the right displays the tables that you selected in the ODBC Data Sources dialog box. The Non- Selected Tables are those that were not selected. Use the Add and Delete buttons to revise your list of Selected Tables and click OK.

- Click OK to continue

Defining relationships graphically

- Select the Graphical View tab. The graphical Database Relationships dialog box will be displayed.

- The fields of each selected table will be displayed. To create a link between the fields, select a field in the first table and drag the cursor to the related field in the next table.

- To create additional relationships, repeat the previous step.

- You can use the command buttons on the right as follows:

    With Access files, you may have *WizSame* define relationships automatically, by clicking Autodetect. The system will build the relationships according to identical field names in the tables *and*

field types. If a relationship between two tables of non-similar names and types has been defined previously, Autodetect will not delete the relationship, and it will not create any new relationship between those tables.

When reading tables directly (*not* through ODBC), *WizSame* will identify relationships defined previously in the database itself.

To revise your selection of tables, click the Tables button. The Tables dialog box will be displayed.

The Selected Tables list at the right displays the tables that you selected in the ODBC Data Sources dialog box. The Non-Selected Tables are those that were not selected. Use the Add and Delete buttons to revise your list of Selected Tables and click OK.

To enlarge the display to full size, click Full Screen size.

To view the tables in the order in which they were chosen, click Arrange.

To remove certain relationships, select them and click Delete.

To remove all defined relationships, click Delete All.

- When you have defined all relationships, click OK.

- To redefine the database relationships at any subsequent point, select File - Database Relationships to display the Database Relationships window.

Reading through ODBC

ODBC (Open Database Connectivity) is a software standard that enables one program to read a data set created by another program. Each database maintains its own specific driver. *WizSame* comes with a set of ODBC drivers that are included in Microsoft's ODBC Software Development Kit.

To work with ODBC data sources, you must first be sure that you have installed the ODBC drivers for the database(s) that you wish to open. Once the drivers are installed, you may define ODBC data sources.

WizSame operates like all standard Windows programs that incorporate ODBC-reading drivers. If *WizSame* fails to open a table through ODBC, first check if another application – such as Microsoft Access – can open it using the same ODBC driver. If that application cannot open the table, you should consult with your dealer concerning your ODBC installation. The Appendix of this manual, "Working with ODBC" may also be of assistance.

- In the Open Data of Type list box select ODBC, or click on the Open Data of Type button if ODBC has been already selected. The Data Sources dialog box will be displayed.

- This dialog box displays all previously defined data sources. A data source is the name given to a connection between a database driver and the database file or directory. Following the Windows standard, this window is divided into two panes:

    The left pane displays the tree of data sources and their tables and views (queries). To display the contents of the current level in the right pane only, click once on the source name or icon. To display the contents of the current level in both the right *and* left panes, double-click on the source. To view the contents of the data source in the *left* frame only, click on the sign at the left of the icon to display its sub-levels.

    The right pane displays the contents of the selected data source.

- Up to three levels can be displayed for a data source: the data source, its tables and the fields of the tables (for display only). Each level is displayed by clicking the sign or double-clicking on the name of the upper level.

- If the data source that you require is displayed, skip to the next step. If you have not yet defined your data source for a file or if you wish to reconfigure an existing data source, click the Data Sources button. A Data Sources dialog box will be displayed.

- Use this dialog box to add a new data source or reconfigure an existing one.

Adding a new data source

Click the Add button in the Data Sources dialog box. The Add Data Source dialog box will be displayed for you to select the data sources you will use with *WizSame*.

From the Installed ODBC Drivers list, select the driver you need and click OK. A driver-specific Setup dialog box will be displayed for you to define the data source.

In the Data Source Name text box, enter a name for the data source, which will be used in *WizSame*. You may add a description of the source in the Description text box. Always provide a unique name – such as Payroll or Accounts Payable – for each source, and provide the specifications required to connect the driver to the database.

After configuring your data source, click OK. The Data Sources dialog box will be displayed again.

Select the required database and click Close to return to the ODBC Data Sources dialog box.

Modifying an existing data source

From the Data Sources (Driver) list, select the data source to be reconfigured and click Setup. The relevant Setup dialog box will be displayed, as in the following example:

Make your modifications in the dialog box and click OK. The Data Sources dialog box will be redisplayed.

Click Close to return to the ODBC Data Source dialog box.

From the Data Sources list on the left, click *once* on the name of the data source that you require. The list of tables within that database will be displayed in the right pane.

Do *not* double-click on the file name to open it, because this will display the tables only in the left pane of the ODBC Data Sources dialog box. They must be displayed in both panes in order to select them.

From the Contents list on the right, click once on the database table(s) to be analyzed.

Do *not* double-click on a table name, because this will display the names of all the fields in the table and will deactivate the OK button for selecting the table.

You may select consecutive tables by highlighting the first name, holding down the Shift key and selecting the last name. Select non-consecutive databases by highlighting the first name, holding down the Ctrl key and selecting the other file names.

Depending on the type of data source, a number of different dialog boxes may be displayed, requesting you to define certain parameters such as database file or password.

Click OK. If you selected a single table, it will be displayed in the *WizSame* work area. If you selected multiple tables, a Database Relationships window will be displayed. Follow the instructions outlined in the previous section – "Combining Several Tables into One *WizSame* Database" – to define the relationships between the tables.

Reading through OLE DB

OLE DB is a set of interfaces that expose data from a variety of sources by using the Component Object Model (COM).

WizSame comes with a set of OLE DB data providers that are included with the Microsoft Data Access Components (MDAC 2.5).

The Microsoft OLE DB data providers allows access to data, such as Microsoft Access, Microsoft SQL Server ,Oracle, DB2 or access to data through ODBC. (OLE DB data provider for DB2 is not included with MDAC 2.5 Kit).

- In the Open Data of Type list box select OLE DB or click Open Data of Type button if OLE DB has already been selected. The Data Link Properties dialog box will be displayed. The OLE DB Provider List lists all the OLE DB providers detected on your hard disk.

- Select an OLE DB provider and click on the Next button. The data link properties may vary depending on your OLE DB provider. (To open the Microsoft Data Link Help click on the Help button). Use the Connection tab of the Data Link Properties dialog box to specify how to connect to your data and click OK.

- The OLE DB Data Sources dialog box will be displayed. The left pane displays the tree of data sources and their tables and views (queries). The first list in the tree includes all tables and views of the database. The other lists include tables and views of several schemas. A schema is a collection of database objects (tables, views) that are owned or have been created by a particular user. To display the contents of the current level in the right pane only, click once on the schema name or icon. The right pane displays the contents of the selected schema. The database object's name contains the table or view name and the schema name. From the Contents list on the right pane, click once on the database object name(s) to be analyzed and click OK.

- If you select a single database object, it will be displayed in the *WizSame* work area. If you select multiple database objects, a Database Relationships window will be displayed. Follow the instructions outlined in the previous section - "Combining several tables into one *WizSame* database" - to define relationships between the database objects.

Reading ASCII files

WizSame includes a unique ASCII file converter that you can use to read a standard text file (with a *.txt extension) created by a word processor, spreadsheet or any other program.
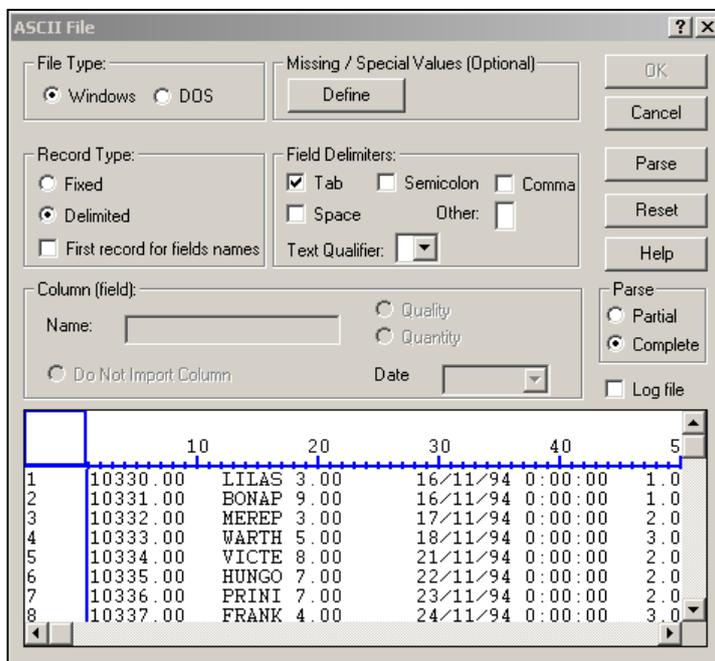
Opening the ASCII File

In the Open Data of Type list box select ASCII, or click on the Open Data of Type button if ASCII has already been selected. The Open dialog box will be displayed.

In the Files of type text box at the bottom, select the ASCII Files (*.txt) option.

Using the Look in list at the top, locate the directory containing the ASCII (*.txt) file. All *.txt files in that directory will be displayed.

Select the file to be analyzed from the list.

Click Open. The ASCII file that you selected will be displayed in the *WizSame* work area, as in the following example:

This dialog box is divided into two main sections:

The top section is used to define parameters for reading the data.

The bottom section is the ASCII file displayed as a table. Any change in parameters in the top section of the dialog box is instantly reflected in the bottom section.

Each line in the ASCII table represents a single record; the number at the left in the first column is therefore the serial number of each record. The numbers in the blue horizontal ruler are the character serial numbers. You may use the horizontal and vertical scroll bars to browse through the table and view all records and characters.

Partial vs. complete parsing

WizSame re-reads the file after each step in order to update the parsing and field types. This may take considerable time when the file is large. You can shorten the time by instructing *WizSame* to refer to the first 1000 records in the file only. The selection between partial and complete parsing is determined in the Parse frame at the right bottom corner of the top section. Select Partial to instruct *WizSame* to read the first 1000 records only, or select Complete to indicate that the whole file should be read after each step.

Converting the file into a readable table

The procedure for converting the ASCII file into a table that *WizSame* can read consists of four main steps:

- Filling in the top section.
- Indicating lines for exclusion.
- Performing field parsing.
- Defining and modifying the field features.

Filling in the top section

- In the File Type block at the top of the dialog box, indicate whether the file was created in DOS or in Windows. This will enable *WizSame* to interpret special characters that are represented differently in each environment.

- In the Record Type block, indicate if the records in the ASCII file are:

  Fixed: All records in the file are of the same length. This is the default.

  Delimited: The records in the file are not of a fixed length. If the data in the table appears to be displayed incorrectly, select this option.

- If you selected Delimited, in the Field Delimiters block at the right, indicate the character by which *WizSame* should delimit the fields:

  Tab, Semicolon, Comma, Space or Other. Once you select the delimiter character, *WizSame* will update the data table, using the selected character to separate the fields in each record.

- If the first line in the file contains the field name, select the First record for field name line.

- In a number of instances, missing values are converted in ASCII files to symbols other than a blank or zero (0), such as an asterisk (*). If you leave these symbols, *WizSame* might read files that were originally numeric as alphanumeric instead. To avoid such a situation, click the Define command button in the Missing/Special Values block at the top of the dialog box. The String/Value Replacement dialog box will be displayed.

- In the String in original ASCII file text box, enter the character that represents a missing value in the ASCII file, and in the Replace with text box, enter a *numeric* value that *does not exist* elsewhere in the data and is quite distant in value from the other numeric values. For example, if an asterisk (*) represents a missing value in the ASCII file, and if all the numeric values are positive, you might enter the value -1000 in the Replace with text box. After you click

OK, the replacement value that you entered will be displayed immediately in the table at the bottom.

- You can use this option to replace other characters as well.

Indicating lines for exclusion

Certain lines in the table – such as subtotals – may contain information that is irrelevant to the data analysis or that interferes with the parsing. These lines must be deleted at this stage.

To indicate lines for exclusion:

- Browse through the table to view all the lines.

- Click on the serial number (in the first column) of each line that you want to exclude from the data set. The text of that line will change color. To cancel your selection, click a second time on the line number.

Performing field parsing

The next step is to define the fields, or perform *parsing*. To assist you, *WizSame* automatically identifies each field in the records and suggests the parsing, in order to create a column that represents a field. The procedure is as follows:

- Click the Parse command button at the top right. *WizSame* will read the records and "reformat" (parse) the table to create and identify fields.

   The black vertical lines separate each field. Above the columns, *WizSame* suggests one of three field types:

   QLT - Qualitative (categorical, alphanumeric characters)

   QNT - Quantitative (numeric data)

   D-M-Y - Date type (date formatted using numbers)

- Check the *WizSame* parsing by scrolling horizontally through the records to view each field. To modify the parsing, you need to add or delete the vertical field separators.

- To *add a field separator,* click the pointer (↑) at the column position of the new separator. Click the body of the table (not in the header area). The new field heading (type) will be displayed as a question mark (?), which means that this field type is yet unknown to *WizSame*.

- To *indicate the field type* of the new column, you can either re-parse or indicate the type in the Column (field) block of the dialog box, as explained in the next step.

- To *delete a field separator,* simply bring the pointer on the separator that you wish to delete, click on the right mouse button, and select delete separator.

- To *move a field separator,* bring the pointer on the separator that you wish to move, and then either press on the left mouse button, and drag the separator while the button is pressed, or click on the right mouse button, and select one of the move commands.

- After creating new fields, you should perform parsing again. To do so, click the Parse command button to update the table. Note that after re-parsing, most question marks are replaced by the field types QLT, QNT or D-M-Y. Fields that remain undefined must be set "manually" in the Column (field) block of the dialog box.

Defining and modifying the field features

To modify the definition of a field, select the column by placing the pointer on the column header and highlighting it.

You may use the Column (field) block to perform three functions on the selected field:

- *Change or define the name of the column (field):* The field name will be printed in the table displayed in the Main dialog box of the *WizSame* work area and later in the *WizSame* reports. The default name for the field is a number, as in field1, field2 and so on. It is

recommended to use meaningful names for each field, as this will facilitate reading the reports later on. Also, you must assign names to each new field, since any added field is unnamed. To name the selected field, type a header for it in the Name text box.

■ *Change or define the field type:* You may:

Change a quantitative field (QNT) into qualitative (QLT)

Change a date field (D-M-Y) into qualitative

Change the date field format

To make your selection, click either the QNT or QLT radios buttons, or pull down the Date list and select D-M-Y format.

■ *Exclude the selected field from the data:* Select the Do Not Import Column radio button.

If you decide to keep the field in the report, simply unclick the column.

Be sure that you have indicated the field type of each column, so that no question marks remain along the top. In addition, be sure that each field has been assigned a name.

To cancel all formatting that you have performed, click the Reset button at the right. The only modifications that will remain are parsing lines.

Once you are satisfied with the parsing and field definitions, click OK. The ASCII file will be displayed as a data set table in the *WizSame* work area.

6. Defining and Issuing the Reports

When you open a data set, it is displayed in the *WizSame* work area, as follows:
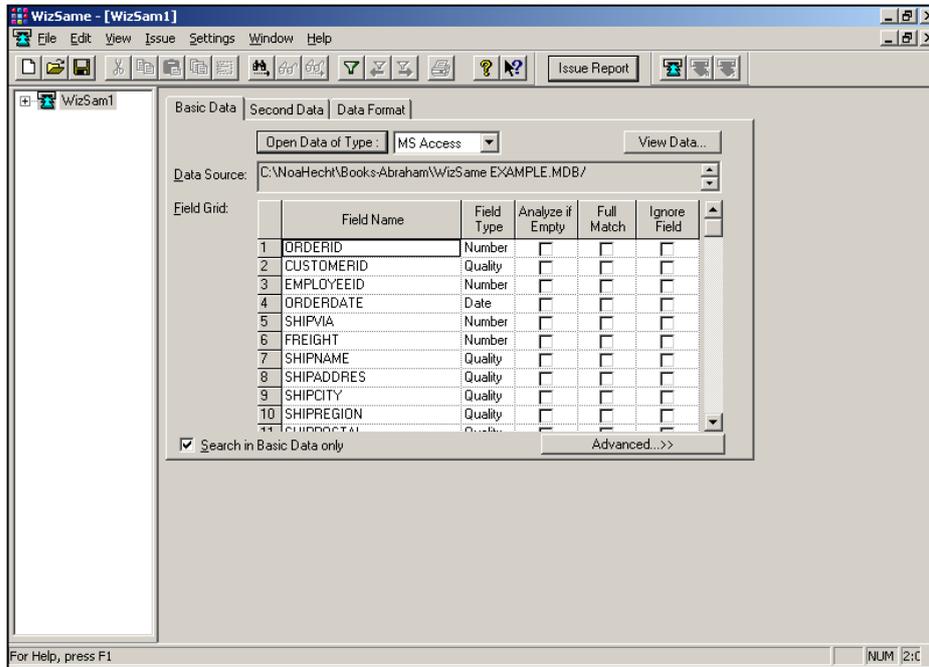
Three tabbed dialog boxes are displayed to assist you in defining the parameters that will be used in analyzing the data and issuing the reports. .

- In the Basic Data dialog box, you define the parameters of the data set fields.

- The Second Data tab should be filled in only when you wish to check a new data set to be merged into your existing data (for example, when you want to reveal existing customers in a new customers list).

- The Data Format dialog box contains settings for the report format, printing, visual design and font selection.

The sequence for issuing the reports is as follows:

- Fill in the above mentioned tabbed dialog boxes.

- Click on the Issue Report button.

- Once the report has been issued, you can review it and export the results.

Basic Data
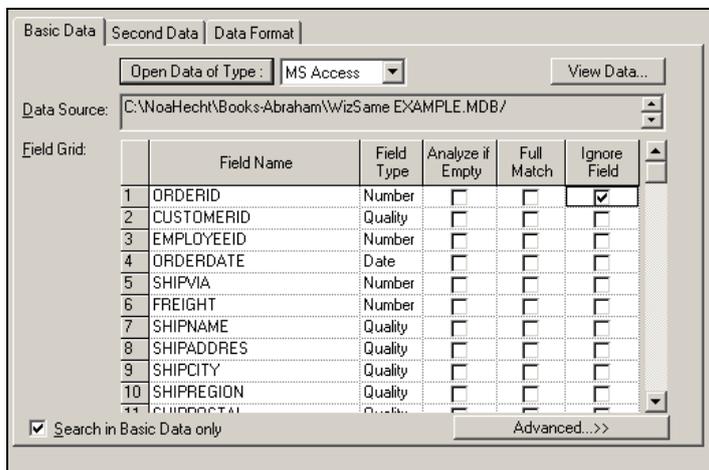
The Basic Data dialog box lists the fields of the data set. Each line refers to a single field.

You can click on the View Data button, if you wish to view the contents of the records. *WizSame* displays the contents of the first 100 records. Click on the arrows on the top of the window to view additional records.

The Field Name is the editable name to be printed in the reports. By default, *WizSame* displays the name as it appears in the data. To edit the field name, click on it and type in a new name.

The Field Type column defines whether the field type is Category, Number or Date

- Quality indicates alphanumeric data; for example, Item Number, Address, Reference Number, Occupation, or Yes/No fields.

- Number indicates numeric data to which mathematical computations can be applied; for example, Total $, Weight, Length or Percent.

- Date is the date data in one of the standard formats (D-M-Y, M-D-Y, etc.).

When *WizSame* reads the data set, it assigns one of these types to each field. When a field contains letters, *WizSame* determines that the field is Quality. When a field contains only numeric data, *WizSame* assumes that the field is a Number. If a field containing only digits is in fact Quality, such as a phone number, the user should change the field type.

If you are not sure whether a certain field should be analyzed as numeric or categorical, you can click on the View Data button, to see the contents of the records in the data.

Select the Match if Empty column to indicate that two records might be presented as duplicate, even if this field contains a value in one record and is empty in the second. This is useful when you wish to reveal duplicate records, even if some of the values are missing. Both numeric and categorical fields are considered empty if they contain no data. Numeric fields that contain 0 (zero) are not considered empty.

Select the Full Match column to indicate that two records are duplicate records, only if their values in this field are *identical*. By default this column is un-checked, meaning that *similarity* of the values is enough for determining duplication.

To exclude a field from the analysis, click in the Ignore Field column.

Example: If you fill the grid as follows –

| Field Name	Full Match	Ignore Field
Name		
Phone #	V	
Email	V	
Address		V

Then the command is: Find all the sets of records, where in each set –

- The values in field Name are similar or identical, *and*
- The values in field Phone # are identical, *and*
- The values in the field Email are identical.

(Ignore the field Address).

Tip: the more fields ignored the greater the number of matching sets that will be discovered.

You can use the Update check boxes option from the Edit menu to update *all* the rows with one command: select or clear all the rows of the Analyze if Empty or Ignore Field columns, or change the field type in all the rows.
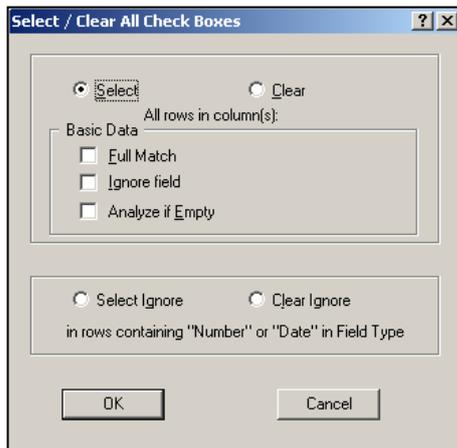
For example, to select all the rows in the Full Match column, select the Select and the Full Match check boxes. To change the Number field type into Category in all the rows, select Replace in the bottom pane. Select Clear to clear all the rows or Undo Replace to revert to the original field type.

Select the Search in Basic Data check box (the default) to analyze one data set. When you search, for example, for duplicate payments in a financial table, this is the right selection. Un-check this check box when you search for duplicates in another table (selected in the Second Data tab). For example, if you analyze a new list of potential customers in order to find customers that already exist in your data, this is the right

selection (when the existing customers are in the Basic Data, and the potential customers are in the Second Data).

Advanced

Select the Advanced button to formulate a command that may contain the OR operator.

In each line you enter a condition: is similar, is equal, ignored, is not similar and is not equal. Select the value in each field.

For example, if you wish to find all the groups of records having similar phone # OR the same email address, select as follows:

Field	Relation	And / Or
Phone #	is similar	Or
Email	is identical	

If you wish to find all the groups of records having similar phone numbers AND not similar email addresses select as follows.

Field	Relation	And / Or
Phone #	is similar	AND
Email	is not similar	

Value conditions

Select the Value conditions button to formulate a command that may contain the OR operator.

Enter the selection criterion and click the OK button. For example, to issue a report where all the matching sets include only records where the dates are between January 1st and March 31st, enter the following values:

Field Name	Operator	Value	And / or
Date	greater than or equal	01/01/2012	and
Date	smaller than or equal	03/31/2012	

Second Data

If you analyze two tables, enter the second table in the Second Data tab. For example, if you analyze a new list of potential customers in order to find customers that already exist in your data, this is the right selection (when the existing customers are in the Basic Data, and the potential customers are in the Second Data).

Make sure to un-check the Search in Basic Data line.

To open the data set in this tab, follow the instructions in the previous chapter for opening a data set.

Data Format

The Data Format dialog box contains settings for the report format, printing, visual design and font selection, which determine the overall appearance in which the data is presented.

Number and Currency

In the Number and Currency Format block, you can change the formats for currency and numbers. System Defined (default) is the default setting in the Regional Settings option of the Control Panel.

To change the format:

Clear the relevant check box (Number and/or Currency).

In the Digits text box, indicate the number of digits to be displayed after the decimal separator.

In the Decimal Separator box, type the character that will be used as the decimal point.

In the 1000 Separator box, type the character that should be used as the thousands separator.

The Example box will illustrate the result of your selections.

Date

In the Date Format block at the right, you can change the format for dates.

Clear the System Defined check box.

In the Format box, select the desired date format from the drop-down list box.

In the Separator box, type in the character that should be used to separate the day, month and year.

The Example box will illustrate the result of your selections.

Heading

In the Subheading text box, you may type a title that will appear at the beginning of the reports.

Font

To change the font type, style and/or size, click the Font command button to display the standard Font dialog box.

Select the Font, Font style and Size and click OK. The reports will display the data in the font and styles selected.

Issuing the report

Once all the parameters are set, click the Issue Report button.

WizSame will display a progress bar where you can see the process stage and the % progress.

The process of finding the matching sets composed of various stages:

The Reading phase, where *WizSame* reads the data and builds tables needed for calculations.

The Calculation phase, where *WizSame* builds the basic tables.

7. Reviewing the Report

The *WizSame* viewer displays the first matching set. The screen is divided into four panes:

The upper two frames display the contents of two matching records.

- A green checkmark ✔ designates identical values.

- A yellow checkmark with a red x ✗ designates similar values.

- No mark is displayed when the field was ignored.

- A red x designates not similar or not identical values.

The frame at the bottom on the right-hand side displays the number of all the records in the current matching set. This information is relevant when there are more than two records in the matching set. The contents of first two records are displayed in the upper frame. To view the contents of another record, double-click on the record serial number. The fields and their values of that record will be displayed in the upper frame on the right-hand side.

Matching sets

The frame at the bottom on the left-hand side lets you scroll between the matching sets and go to certain record.

The first line displays the serial number of the current matching set, and the total number of the matching sets discovered in the data. To go to the next or previous matching set, click on the arrows. To go directly to a certain matching set, enter the matching set number.

The second line allows searching for a certain record. To find if a certain record is included in a matching set, enter the record number.

> Tip: when the matching criterion includes several conditions connected by the OR operator, a certain record may be included in several matching sets. *WizSame* displays the first.

Printing and exporting the report of matching sets

To print the report on the printer or export the report to another file select the print option . *WizSame* will display the Print matching sets dialog box.

In the Print to box select either Printer to print the report on the printer; ASCII or RTF to create a text file of the report; or MS Access to export the matching sets to a database.

In the Print / Export Range box select either All to print (or export) all the matching sets, or enter a range of serial numbers in the SET NUMBER from… to… box to print (or export) some of the matching sets.

You can enter any text in the Heading box, which will be printed as the title of the report.

Print matching sets

The Issue menu includes the option Print Matching Sets. Select this option to print all the discovered matching sets. WizSame will save all the duplicate records in a text file adding two new columns stating the number of the matching set and the data set number. Please note when only one file is analyzed the data set number will be 1. For two data sets the number will be 1 or 2. Note that when you open the issue menu after selecting a group of matching sets (using the filter option) the name of this command is converted to Print Selected Sets.

Selecting a group of matching sets

If *WizSame* found a large number of matching sets, you may wish to select a group out of these sets. This may be useful when you want to concentrate on a certain group. For example, if *WizSame* found, say, 1,000 matching sets of duplicate payments, you may only wish to select the sets from the last three months.

To select a group of matching sets –

Click on the Select icon. *WizSame* will display the following window:

Enter the selection criterion and click the Issue button. For example, to issue a report of the matching sets, where the dates are between January 1st and March 31st, enter as follows:

Field Name	Operator	Value	And / or
Date	greater than or equal	01/01/03	and
Date	smaller than or equal	03/31/03	

Note that a matching set meets these conditions if the condition holds for at least one of the records in the set. For example, if a matching set

includes two records, and the date of one is before 01/01/03 while the date of the other is after 01/01/03, this matching set meets the first condition presented above.

WizSame displays the selected matching sets in the Selected Set report. To open the report, select it from the tree menu.

This report is very similar to the Matching Sets report. The only differences are:

- The report lists just those matching sets that meet the selection criterion.

- The upper frame displays the selection criterion.

- The fields relevant to selection criterion are marked by a gray background for the checkmark in the column on the left-hand side - .
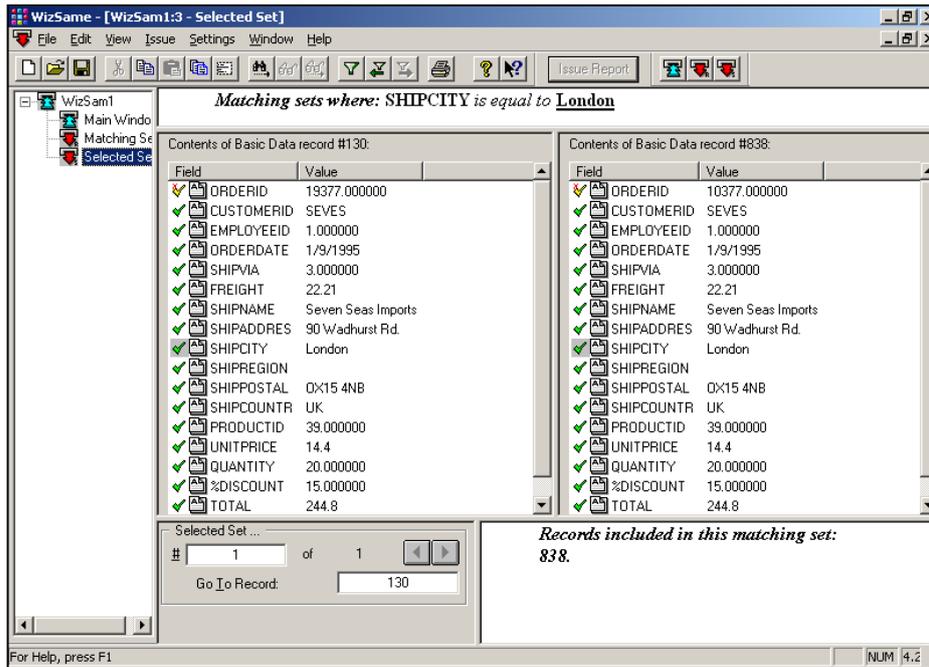
When reviewing the Selected Set report you may re-issue a selection. Select the **Selection** button. *WizSame* will display above window, where you can enter additional selection conditions. The conditions will be applied on the previously selected group (as if the old and the new conditions were entered together). For example, after issuing a matching report containing records where the date is between 01/01/03 and 03/31/03, you can now add the condition that the State be either NY or Texas. *WizSame* will apply this condition on the previous report, and so the records in the new report will meet both conditions:

- The date is between 01/01/03 and 03/31/03, and

- State is either NY or Texas,

You can continue adding conditions with no limit.

Creating a new file without duplicate records

When you search duplicate records in a new list (that is records in the Second Data that are identical or similar to records you already have in the Basic Data), *WizSame* lets you create a new file without the duplicate records. For example, if you check a new customers list (opened in the Second Data tab) in order to find existing customers (opened in the Basic Data tab), then after issuing the matching sets report *WizSame* lets you create a new file that contains just the new customers (that is, without the duplicate customers). The new file created by *WizSame* is an ASCII (text) file.

To create the file select Print to a new file option from the Issue menu. *WizSame* will display the following window:

This window lets you determine which records will be included in the new file:

Select the Print records in matching sets (duplicate) option, to create a file that contains just the duplicate records.

Select the Print records not in matching sets (non-duplicate) option, to create a file that contains all the records, except the duplicate records (that is, the records included in matching sets).

Select the All option, to create a file that contains both duplicate and non-duplicate records.

Note: When you are analyzing two sets of data the printed records are from the data that was selected on the Second Data tab.

If you selected a group of matching sets (see previous section), WizSame displays a window that includes the following additional option:

Select the Print records in selected matching sets option, to create a file that contains just the *selected* duplicate records.
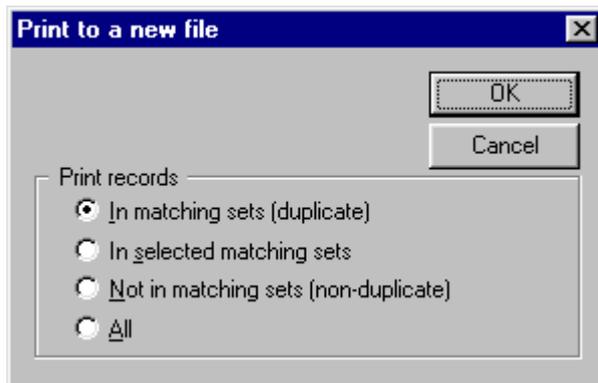
Updating the Synonym dictionary

WizSame includes a Synonym dictionary. For example, one of the synonyms is:

New-York, New York, NY, N.Y.

When one record contains one of these values in a certain field, and another record contains another of these values in the same field, the two records are considered identical in regard to this field.

You can update the Synonym dictionary. You may

- Add new synonyms
- Edit existing synonyms.

The dictionary is saved in the WizSyn.txt file. This file is saved in the *WizSame* program directory.

You may open the file and edit it using Excel. Each line contains a list of synonymous names. You may insert new names, delete or replace existing names. Each name should be inserted in another cell.

When saving the file, make sure you save it as a text file - select Tab delimited from the Type list.

8. Frequently Asked Questions

Q: What is the largest size of data set that WizSame can practically analyze?

A: The size of the data set is neither limited by the number of fields nor the number of records.

Q: How can I decrease or increase the number of matching sets in the report?

A: The following parameters determine the number of matching sets:

- Ignore Field: When you select to ignore a field, *WizSame* may reveal more matching sets. For example, if you ignore all the fields except for the City, then every group of records having the same city constitutes a matching set.

- Full Match: When you check the full match column, WizSame may reveal less matching sets. As long as this column is unchecked, similarity of values is enough for establishing a matching set.

- Conditions connected by OR: When you determine the matching criterion in the Advanced window, you may increase the number of matching set by connecting the conditions OR.

Q: How do I increase the speed of issuing the report?

A: The main methods include:

Deleting uninformative fields by selecting Ignore in the field grid.

Increasing the RAM of your computer. (Note that if not enough RAM is available, it causes *WizSame* to use the hard drive resources to issue the calculations, decreasing its efficiency by a factor of 100).

Q: How do I save the report?

A: Whenever you operate *WizSame*, it creates a *.wzs file, which includes the parameters of the analysis and the last issued reports. You can save this file through the File - Save or File - Save As option. To open the saved file, select it through the File - Open option (or from the list of last four saved files).

Note that when you reopen a file, *WizSame* keeps the previously issued report until the new ones are actually issued; then it replaces the report with the new saved report. If you want to save the previous report, use the File - Save As option to save the *.wzs file under a different name, and reissue the analysis from the beginning.

Q: What is the structure of the Microsoft Access table created by the WizSame export option?

A: *WizSame* exports the matching sets to Microsoft Access in the following two formats:

Spreadsheet format: Each record is written in one line, where each column refers to another field in the data set. The first column, MatchNum denotes the matching set number. All the records (i.e., lines) having the same matching set number belong to the same matching set.

One field in a line: Each field is written in another record (similar to the printed format).

Both formats appear under *queries* in the Access database.

Q: How do I create a table of all of the duplicate records?

A: Use the Print Matching Sets from the Issue menu to create an ASCII file where each record is in one line. The first column denotes the matching set number, the second column denotes the record number and the third column denotes the data set number.

Q: How do I instruct WizSame to ignore certain words (such as street, road, ave., etc.)?

A: In the *WizSame* synonym dictionary, enter a line where the first cell is blank, and the other cells contain the words to be ignored (street, road, etc.). This line tells *WizSame* that blank is synonymous with these words, and therefore they are ignored.

Q: How do I save the non-duplicate records while I am analyzing one data set?

A: Use Print to a New File from the Issue Menu, choose the Print Records Not in Matching Sets option to create a new ASCII file that contains all the records except the duplicate records.

Q: How do I instruct WizSame to reveal duplicate records in my Vendor data and Employee data where the name is similar and the address is not similar?

A: After opening the data sets click the Advanced button in the Basic Data Tab. In the Field Name column click on the first empty field and select address. In the Similarity column select "is not similar" and in the And/Or column select "and". Click OK then click on the Issue Rules button.

Q: How do I change the screen and the font size?

A: In the Windows menu select Zoom and then select the screen size (small, medium, large) or the Set display font.

Quick Reference Guide

After you have familiarized yourself with *WizSame*, you may wish to use this chapter to refresh your memory about how to perform certain *WizSame* processes or how to use a certain function available through the *WizSame* menus.

The *File* menu

This menu provides options for performing global operations on data files.

New opens a new *WizSame* working file. This enables you to select a new data set for issuing rules.

Open opens a previously saved *WizSame* working file (*.wzs).

Close closes the current *WizSame* working file.

Save saves the current working file

Save As saves the current *WizSame* working file with a new name.

Open Basic Data opens a data set for analysis. The data type may be Text, dBase, MS Access, MS SQL, Oracle, ODBC or OLE DB.

Open Second Data opens a second data set (when you want to reveal in a new data set records that are similar to the records in the basic data).

Data Settings enables you to edit the database connections (if several tables are selected).

Print prints the current report on the printer or exports it as an ASCII file or an MS Access table.

Exit closes the *WizSame* application.

The *Edit* menu

The contents of the Edit menu change according to the information currently displayed in the *WizSame* screen. This section describes special Edit functions.

Update check boxes enables you to select / clear all the fields in the Analyze if Empty, Full Match or Ignore Field column.

The *Issue* menu

This menu is used to generate reports and to select matching sets.

Print Matching Sets issues a report of all the matching sets.

Select Matching Set... opens a window were you can enter conditions for selecting matching sets and saving them in another table.

Print to a new file creates a file that contains all of the duplicate or the non-duplicate records or both.

Index