

SOFTWARE FEATURES AND FUNCTIONS

Linkage Software Details	
LinkageWiz Data Matching Software	
Distributor	LinkageWiz Software 24 Albert Street Payneham South Australia 5070 Australia Phone : +61 405 844 024 Email : linkagewiz@internode.on.net
Other Information:	Commercial

IT Infrastructure on which the software operates:

- Platforms / Operating Systems

Options: Windows 2000, XP, 2003 Server, Vista
Installed on Windows Server for this evaluation

- Other software required (database packages or editors):

None

- Minimum Hardware Specification:

Recommended: PC with 2Gb RAM, 4 Gb free hard disk space or higher.

- Other Information:
-

Overview of Linkage Software

Description of the software:

- How easy was the system to install?
(Please note specialist support required)

The system was very easy to install.
(No specialist support required)

- Ready to perform linkages “out of the box”? Y N

LinkageWiz is fully operational after installation; the system includes a graphical user interface (GUI) which allows users to configure projects easily.

- Part of a larger system performing other functions (e.g. data integration and cleaning functions?) Y N

If yes, please note functions available

Data Matching, De-duplication and Data Cleansing tool

- Is the software a set of components, requiring that a system be built around them? Y N

LinkageWiz has a number of functions which can easily be configured and operated from the graphical user interface (GUI). These component include:

- Data linkage of records for individuals or companies;
 - Identification and removal of duplicate records for individuals or companies;
 - Identification of data quality issues;
 - Data standardization and cleansing, including address standardization.
-

- Is the software well documented? Y N

Comprehensive software manual available.

List of publications available on LinkageWiz website:

<http://www.linkagewiz.com/Research%20Papers.htm>

Software features:

- Can the system schedule linkages? Y N

Not without adding to the existing software (although it may be possible to add to the existing system)

- Options to use reference files and or lookup tables (e.g., geocoding). Y N

LinkageWiz contains a number of lookup tables that are used for data matching and data cleaning:

- Nick names library
- Phonetic non-matches
- Excluded values from matching process
- Remove organizational noise from business names
- Property types
- Street types - spelling variations and standardized abbreviations
- Street suffixes - spelling variations and standardized abbreviations
- Postal boxes - spelling variations and standardized abbreviations
- City names - spelling variations and standardized abbreviations
- States and provinces - spelling variations and standardized abbreviations
- Countries - spelling variations and standardized abbreviations
- First names and most common associated sex

LinkageWiz also includes a geocoding engine (using combinations of exact and fuzzy logic) that allows for the geocoding of address data. Latitudes and Longitude coordinates can be assigned to data using (user defined) geographical reference files.

- Does the software allow multiple users? Y N

If yes, does the software manage the integrity of data files and prevent data corruption when multiple users access the system? Y N

Although as it is a database structure it may be possible to add to the existing system.

- Does the software contain security protection, access and authentication features? Y N

No authentication access features.

- What is the maximum file size (number of records) that the software can handle (Please indicate any file size limitations)?

Maximum number of records per table: Approximately 4-5 million records (Enterprise version required).

Files with more than 4-5 million records can be processed using a special Chunking methodology.

Maximum Database Size: 2 Gigabytes.

- Does the software require any parameter files? Y N

Data Dictionary Y N

Weight Files Y N

Please note any others?

LinkageWiz contains a number of lookup tables that can be used for data matching and data cleaning:

- Nick names library
- Phonetic non-matches
- Excluded values from matching process
- Remove organizational noise from business names
- Property types
- Street types - spelling variations and standardized abbreviations
- Street suffixes - spelling variations and standardized abbreviations
- Postal boxes - spelling variations and standardized abbreviations
- City names - spelling variations and standardized abbreviations
- States and provinces - spelling variations and standardized abbreviations
- Countries - spelling variations and standardized abbreviations
- First names and most common associated sex

Note: although there are no match weight files, a set of default linkage weights have been included with LinkageWiz. These have been pre-configured but can also be modified by the user.

Pre Processing (ETL)

Standardisation components:

- Does the software include a Standardisation Component? **Y** **N**

LinkageWiz prepares the data prior to the commencement of the record linkage process to reduce minimize the impact of spelling, typographical and other data entry errors upon the accuracy of the linkage:

- Case changes (Lower, Upper and Proper);
- Removes extraneous noise and punctuation;
- Full name standardization;
- Business name standardization;
- Standardizes street address information;
- E-mail standardization;
- Computes date of birth components;
- Creates phonetic codes for name fields (*NYSIIS* for family name and *SOUNDEX* for given names).

Recoding - A Field Recoder module allows for custom recoding schemes to be developed and saved for future use, useful when multiple files need to be recoded consistently or if the same file needs to be recoded on a regular basis.

Truncation - function allows for the truncation of the contents of an existing field. The results may be saved into the same field or optionally into a new field. Contents may be truncated a specified number of characters from either the left or right.

Sex Derivation - function allows the derivation of a new field containing the likely sex of the client based upon information supplied in the first name or given names fields.

Concatenation - function allows two fields to be joined together to create a new field.

If Yes.....

- Can standardisation/parsing components be customised (e.g. name and address)? **Y** **N**

Standardisation components can be changed and adjusted for different datasets which will have different values and formats.

- Does any address standardisation conform to Australian standards? **Y** **N**

Address standardization converts street addresses to a standardized format. Spelling mistakes are corrected, superfluous punctuation is removed, capitalization is standardized and address components are translated into standard abbreviations. A standardized address field can be output along with key elements of that address (building number, street name, street

name and street type).

Geocoding data is loaded from external files which can be implemented from Australian standards.

- Can different processes be applied to different files with a linkage strategy? **Y** **N**

It is possible to initialize more than one component standardiser of the same type operating on different fields of a data set. For example, if a hospital data set contains a date-of-birth and a date-of admission field one data standardiser each may be required.

- Can the user choose to append standardised fields to the original data record? **Y** **N**

The cleaned data is stored in LinkageWiz system fields and does not overwrite the original data. Where the system recodes or translates existing data fields, LinkageWiz gives the user the option whether to overwrite or create a new field.

- Can the user take subsets of the data fields for processing? **Y** **N**

System allows the user to export a subset of processed fields.

Data Management:

- In what file formats can the software use data?

Flat Files Y N

LinkageWiz can import delimited and fixed-width text files. Delimited files contain the field values separated by commas, spaces or other characters; in a fixed-width file each field has a defined width.

Text files sourced from UNIX systems cannot be imported directly and need to be converted using the LinkageWiz UNIX conversion utility.

Analytical dataset Y N (For example SPSS, SAS, STAT etc)

If Yes, please list.....

Microsoft Excel

SPSS (by exporting to DBASE format or via the ODBC data import facility)

Database Y N

If Yes.....

What types of database can be employed?

Microsoft FoxPro (using the ODBC data import facility)

Paradox

DBASE

Enterprise wide databases such as SQL Server, Oracle and others (using the ODBC data import facility)

Does it take advantage of database indexing?

Yes – During the preparation for matching indexes are applied based on the defined matching variables.

-
- Does the software contain a utility for viewing data records? Y N

Viewing Data Tables - This option allows you to view a data table that has been imported. A grid containing the contents of the table will be displayed.

-
- Does the software permit the user to partition/concatenate variables to maximise the use of the information they contain? Y N

Recoding – system allows custom recoding schemes to be developed and saved for future use.

Truncation - system allows the truncation of the contents of an existing field. Contents may be truncated a specified number of characters from either the left or right.

Concatenation - allows two fields to be joined together to create a new field.

Record Linkage Methodology

Types of linkage supported:

- Deduplication/Internal Linkage i.e. one file linked to itself. **Y** **N**

Scanning for duplicate records within a single table:

- Choose only a Source table if you are running a duplicates analysis
-

- Linkage two files i.e. linking files to the Population Spine. **Y** **N**

Matching records from two separate tables:

- Choose both Source and Reference tables if you are undertaking data matching.
-

Linkage passes and blocking strategies:

- Does the software support multiple linkage passes (with different blocking and different linkage variables)? **Y** **N**

Multiple passes through the data can be used; each for a separate blocking variable.

- How does the software block variables?

Blocking subdivides data into a set of mutually exclusive subsets under the assumption that no matches occur across different blocks.

- Do users set blocking variables? **Y** **N**

Linkage *Wiz* allows the user to select from the following blocking variables:

- Family Name
 - Business/Company Name
 - Street Address
 - Date of Birth
 - Date of Death or other date (eg service date)
 - Australian or Nova Scotia Medicare Card Number (or other number where the 1st 9 digits are unique to a given family)
 - Unique Identifier (eg Social Security Card Number, Veteran Card Number)
-

- Can the user specify how the system handles missing values for blocking **Y** **N**
-

Linkage weights:

- Does the system compute matching weights? Y N
(Please describe how these weight are calculated)

LinkageWiz uses *Probabilistic Record Linkage* techniques to link together records for the same individual. The sum of the weights across all demographic characteristics is used to derive a *Total Score*. Depending upon the magnitude of the total score, record pairs are considered to be linkages, potential linkages (that possibly require manual inspection) or non-linkages.

The linkage weights are higher for variables with more specificity (such as family name), and lower for variables with less specificity (such as sex) The weights are calculated from the logarithm (base 2) of the frequency ratio of the field being examined (Binit Weights).

Once the record pairs have been generated LinkageWiz merges all record pairs that belong to the same individuals to create *Linkage Groups*.

- Does the system allow manual manipulation of matching weights? Y N

A set of default linkage weights has been included with LinkageWiz. These have been pre-configured for most types of datasets.

Note: These weights can also be modified by the user.

- Does the system allow agreement and disagreement weights? Y N

The options allow the user to specify an agreement weight or to specify both agreement and disagreement weights.

The demographic characteristics between a pair of records are compared, with field agreements assigned a positive linkage weight and disagreements assigned a zero or negative linkage weight. Invalid data values, default values or missing data are excluded from the assignment of weights.

Linkage Functions:

- Does the user specify linking variables? **Y** **N**

The user can choose from the following linkage variables:

Client Number
Business Names
Family Name
Previous Family Name
Given Names
First Name
Middle Name / Initial
First Name Alias
Gender
Date of Birth
Date of Death
Street Address
Postcode
Medicare Number
Unique identifier
Diagnose(s)
User Defined 1
User Defined 2
User Defined 3
User Defined 4
User Defined 5

- Does the user specify the types of comparisons? **Y** **N**

A set of comparisons (optimised by field type) have been included in LinkageWiz – with additional User Defined matching variables.

- Is there a maximum number of linking variables? **Y** **N**
If Yes, how many?

21 Linkage variables (see specify linkage variables above)

- What kinds of comparison functions are available for different types of variables?

Field specific comparison functions are applied during matching (i.e. comparison functions are tailored to field types)

- Can the user specify how the system handles missing values for comparisons? **Y** **N**
-

- Does software contain/support routines for estimating linkage errors? **Y** **N**
-

- Can the user specify critical variables that must agree for a link to take place? **Y** **N**
-

- Can the user specify subsets of the data to be linked? **Y** **N**

Match Data/Scan Dups allows the user to select all record or to process a 1% or 10% sample of records/

Post-linkage Functions

Clerical Review:

- Does the software contain a utility for review of possible links? **Y** **N**

If Yes, does the system allow multiple reviewing users? **Y** **N**

If Yes, does it allow examination of match weights for each variable? **Y** **N**

LinkageWiz allows users to inspect the individual agreement and disagreement weights assigned to an individual record pair.

If Yes, what kind of functionality is provided?

Pairs **Y** **N**

Groups **Y** **N**

Matched record pairs can be examined in the *Matched Pairs Analyzer* to assist in the determination of linkage thresholds and to scan for *False Positives* (invalid matches) and *False Negatives* (matches that have been missed).

In the Matches Report all related records for an entity are automatically grouped together. Each group of records is assigned a unique identifier named the Group ID. The Match Scores and Match Variables are also displayed in the report columns.

-
- Can pairs/groups of records be updated? **Y** **N**

If Yes, Can the user “back up” or restore possible links before committing to decisions?

Matched Pairs Analyzer - Invalid record pairs can be flagged for inclusion/exclusion from the results. Once the record pairs have been edited the Linkage Groups will be automatically recreated.

Linkage Groups can be edited in the *Clerical Review Screen*. Records can be flagged for inclusion/ exclusion and be re-assigned to other groups or to new groups as appropriate. As Linkage Groups often contain several records there is often more evidence to support whether a record belongs to the same individual than in the *Matched Pairs Analyzer*.

-
- Can software recomputed match weights based on clerical decisions? **Y** **N**

Output manipulation:

- Can the user customise the outputs? Y N
If yes, please describe.

Customisation and manipulation available on outputs.

- Can a “master” record be created that combines values from two or more records for different fields Y N

Not Automatically.

- Does the software contain a utility that extracts files of linked and unlinked records? Can the user specify the format of such extracts? Y N

Export data after data cleaning to the following export destinations and file formats:

- Current/New LinkageWiz project
- MS Access database
- MS Excel
- Comma Delimited Text File

Export data after data matching:

Deduplicate (destination and formats as above)

- All Records - Exports all records
- Duplicates Only - Exports duplicate records only
- Unique Records Only - Exports unique records only
- Matched Pairs - Exports Matched Pairs

One to Many (destination and formats as above)

- All Matches - Exports the matches with the records grouped for each entity
- Non-Matches (Source) - Exports cases from the source table that have not been matched.
- Non-Matches (Reference) - Exports cases from the reference table that have not been matched
- Matched Pairs - Exports Matched Pairs

Matched Pairs can be exported to a wider variety of file formats.

Linkage performance (statistics):

- What linkage performance (Statistics) outputs are available?

Match Scores Histogram displays a graph of the frequency profile from the last linkage run.

Log information displayed through dialog boxes provide the number of record pairs compared and how long it took to compare them.

- Does the software contain a utility for generating reports on the linked, unlinked, duplicate, and possible linked records (can it be customised and printed)? **Y** **N**

Matched Pairs Analyzer - allows individual Match Pairs to be reviewed and edited. Each match pair is displayed in separate rows, with match fields for the source and reference table displayed in alternating columns.

In the Matches Report all related records for an entity are automatically grouped together. Each group of records (known as a Linkage Group) is assigned a unique identifier named the Group ID (refer Group ID report column). The Match Scores (labeled Confidence) and Match Variables are also displayed in the report columns.

Matched Pairs and the Matches report can be customised and can be exported to a wider variety of file formats.

Summary

- Is the software reliable?
Any system crashes during evaluation? Y N
Replicable Results across runs? Y N
-

- Have you developed any custom solutions you would be willing to share?

Not Applicable

- Do you know of additional features are planned with the next version of the software (please describe).

Not known

- Please indicate any performance issues during evaluation.
-

- Any Other Information:
-