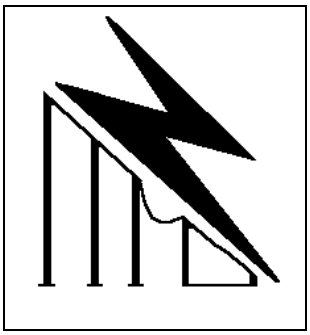
NATIONAL LOAD RESEARCH (LR) PROJECT: SITE REFERENCE DATABASE DESCRIPTION

Document Ref : 20050803 SiteRef database description - V2.doc



J Cheek

Marcus Dekenah Consulting cc
+27-12-667-1347

25 Maluti Ave
Doringkloof
0157

Table of contents

1. INTRODUCTION

Over a period of some years, Marcus Dekenah Consulting (MDC) has conducted Load Research at several electrified sites in South Africa. As part of this research, consumption logging devices have been installed at some of these sites, and teams of field researchers visit the monitored dwellings annually to interview the occupants and to gather socio-demographic data relating to their electrical appliances and usage patterns. In addition to socio-demographic data, the researchers gather data relating to the dwelling and the associated electricity network (eg addresses, GPS positions, etc). Collectively, this data constitutes a 'Site Reference'. In addition to the annual Site Reference, surveys are conducted when logging devices are commissioned at a new site. These are called 'Installation References'.

Over time, the combination of normal effects, collection errors, transcription errors, and differences in data collection techniques make it difficult to compare year-on-year data for the same dwelling. It has thus become necessary to normalize and combine the data resulting from these References into a single consistent database that can be easily augmented as more data is added in years to come.

Broadly speaking, Installation and Site References exist for the years 2000 to the present. This document describes how data for these years has been normalized and combined into a single Microsoft Access 2003 database. It contains reference information for that database, as well as guidelines for future upkeep and maintenance. Note that development of the database is ongoing, both in terms of adding new data (eg new survey References) as well as developing more sophisticated query tools to improve the data quality.

2. DOCUMENT VERSION LOG

Version	Date	Description of changes
1	02/08/2005	First draft
2	13/08/2005	Updated to include 'Future development' section, other minor changes.

3. REFERENCE DOCUMENTS

None.

4. ABBREVIATIONS

Abbreviation	Full description
DB	Database
DD	Decimal degrees (eg 10.12345°)
DDM	Degrees, Decimal Minutes (eg 10 32.12345 =10° 32.1235')
DMDS	Degrees, Minutes, Decimal Seconds (eg 10 32 45.1 = 10° 32' 45.1")
GPS	Global Positioning System
IR	Installation Reference
LR	Load Research
PR	Partial Reference
SD	Socio-Demographic

Abbreviation	Full description
SR	Site Reference
STS	Standard Transfer Specification
WGS84	World Geodetic System 1984

5. SUMMARY OF CURRENT DATABASE FORMAT AND STATE

At this time, data pertaining to each monitored dwelling has been captured using primary fields similar to those specified in the following table :

Field Name	Data Type	Description
Site	Text	The name of the site where the reference was taken
LoggerID	Text	The ID of the logger at the LP
ChannelNo	Text	The channel number of the logger at the LP
Lat	Number	GPS Latitude of the LP-monitored dwelling, wrt WGS84 and in the format dd.ddddd
Lon	Number	GPS Longitude of the LP-monitored dwelling, wrt WGS84 and in the format dd.ddddd
Erf	Text	Erf number of the ground on which the LP-monitored dwelling is situated
Eskom	Text	ESKOM number of the LP-monitored dwelling
MeterNo	Text	Meter Number of the LP-monitored dwelling
HolderName	Text	The name (either house head or respondent) of the dwelling occupant
SurvType	Text	Type of survey, ie Installation Refererence, Site Reference or Market Research
SurvDate	Date/Time	The date on which this reference was taken
Comments	Text	Comments relevant to this survey
SheetNo	Text	Number of the sheet within this site reference on which this LP-monitored dwelling's details are recorded
TestType	Text	Character indicating type of test : A=Load at house (surest), B=Load at pole (only in 2001 - rare); C=wiring trace (least certain)
Radius	Number	Distance of dwelling from area centroid in metres

In addition, data pertaining to each of the data loggers used to monitor these dwellings has been captured using a very similar table structure as specified below :

Field Name	Data Type	Description
Site	Text	The name of the site where this Logger reference was taken
Lat	Number	GPS Latitude of the LP-monitored dwelling, wrt WGS84 and in the format dd.ddddd
Lon	Number	GPS Longitude of the LP-monitored dwelling, wrt WGS84 and in the format dd.ddddd
PoleNo	Text	The number of the pole (or kiosk) at which the logger is located
LoggerID	Text	The ID of the logger
NumChans	Number	
SurvType	Text	Type of survey, ie Installation Refererence, Site Reference or Market Research
SurvDate	Date/Time	The date on which this reference was taken
SheetNo	Text	Number of the sheet within this site reference on which this logger's details are recorded
Comments	Text	Comments relevant to this survey
Radius	Number	Distance of dwelling from area centroid in metres

The table below shows, on a per-site basis, the state of the SR normalization project as at the time of writing, using the following summary parameters :

NumLoggers The number of loggers represented by survey data from that site
TotLogRefs The total number of times that all of the loggers at the site have been surveyed
TotPeakChans The maximum number of channels represented by survey data from that site
TotChanRefs The total number of times that all of the channels at the site have been surveyed

JB3 Logger, Logger reference and Channel reference counts per site

Site	NumLoggers	TotLogRefs	TotPeakChans	TotChanRefs
Antioch	32	125	93	371
Dinokana	19	57	81	243
Driekoppies	25	75	82	246

JB3 Logger, Logger reference and Channel reference counts per site				
Site	NumLoggers	TotLogRefs	TotPeakChans	TotChanRefs
Garagapola	34	124	79	290
Gasese	31	124	87	348
Ikgomotseng	24	47	92	190
Khayalitsha	14	42	89	267
Maconqo	22	43	65	129
Mafefe	33	64	87	174
Matshana	29	145	66	331
Mfazazane	29	87	73	220
Peacetown	24	48	87	173
Tambo	34	125	100	397
Vlaklaagte	20	60	90	270

By way of illustrative example :

- At Matshana, the 29 loggers represented by the survey data have been referenced a total of 145 times.
- Historically, these 29 loggers have monitored (or are monitoring) a total of 66 channels, which in turn have been referenced a total of 331 times.
- The two figures suggest that this site as a whole has been referenced approximately 5 times (145 / 29 = 5.00 and 331 / 66 = 5.02)

As at the time of writing, normalized Reference data is available for a total of 370 loggers and 1171 channels. A total of 1166 logger references and 3649 channel references are included in the dataset.

6. DATA CONDITIONING PROCESS

6.1 OVERVIEW

The output from the Installation and annual Site Reference Surveys is generally supplied in the form of an Excel workbook with worksheets captured from handwritten survey documentation. The workbook usually contains at least the following sheets :

'Master' Information relating to the location, identification and configuration of a data logger
'Detail' Information relating to the dwelling monitored by each channel of the logger

The process followed in combining and normalizing the Installation and Site Reference data was as follows :

- 'Master' and 'Detail' worksheets for Site references (and Installation references, where they exist) are linked to tables in the SR DB. This link is not a static (ie one-time) import, but dynamic; ie the database is linked to a "tree" of Excel worksheets, and changes made in the underlying worksheets are immediately reflected in the corresponding database tables.
- A set of 'Presentation' queries dynamically conditions the table generated from each Reference worksheet into a standard format.
- Successive Reference tables for each area are then combined into a single, multi-year table for each site.
- The multi-year Reference tables are then combined into a single multi-site, multi-year reference table. This table contains all survey references.
- In a first stage of processing, various queries are used to analyze the underlying data and

highlight inconsistencies caused by simple errors in the underlying reference worksheets. Where possible, these errors are corrected in the underlying worksheets. An iterative process is followed until the underlying dataset is as consistent as possible.

- In a second stage of processing, more sophisticated queries are used to suggest 'missing' data by interpolation between known data points. The interpolation is conservative and does not synthesize data unless its accuracy is known with a high degree of confidence.
- The outputs from the database are various static tables and query dynasets which contain normalized Reference data. Typically, it is these output tables that are used as input data for other purposes, eg to direct site survey staff.

At the present time, the "more sophisticated" queries referred to above have only been used to augment the dataset used to direct operational staff performing the 2005 Site Reference surveys. Most such queries are in a developmental stateOther

6.2 LINKING AND MANIPULATION OF SPREADSHEETS

With regard to the above process, the following points are noteworthy :

- The link between the Excel worksheet and the associated Access table is dynamic, ie editing done in one application (either Excel or Access) affects data in the other application. However, as a general rule, the dynamic link is only exercised in one direction; viz any changes are made directly in the underlying Excel sheets.
- As a general rule, as little change as possible is made to the intrinsic content of the original Excel worksheet. Where it is necessary to make a change the worksheet structure to enable linking, a copy of the worksheet is made and the changes made in the copy For this reason, many workbooks will have multiple similar sheets, eg 'Detail' and 'Detail(2)'.
- As a general rule, data in the underlying Excel sheets is only edited to correct 'transcription' type errors, not to augment the data contained in it.
- Where changes are made to the cell content, comment text has been pasted into the 'Comment' field in addition to any original comment which may have been present. These post-survey audit comments are identified with the writer's initials in square brackets, eg "Meter No corrected from 041012**5**4135 [JC]".

Examples of the types of errors that have been encountered and corrected in this way are as follows :

- Reducing the extent of spreadsheet title row structures from several rows to a single row, to form a new header suitable for generating field names.
- Deleting apparently blank trailing rows at the bottom of a worksheet to suppress blank records in the corresponding linked table.
- Re-formatting single cells or an entire column as type 'Text' rather than type 'Number' (or vice-versa) to facilitate linking.
- Deleting inappropriate values from worksheet cells, eg where the 'Meter Number' field has been populated with the text "No #" to indicate that no number could be retrieved, the "No #" text would be cut from the Meter Number field and pasted into the 'Comment' field to leave a 'Null' value in the Meter Number field.
- Replacing the content of an entire column of GPS coordinates where position formats have been incorrectly transcribed, eg where DMDS positions have been incorrectly recorded as DDM.
- Swapping entire columns of GPS coordinates where Latitude and Longitude were reversed.
- Correcting simple errors where the value(s) concerned were correctly recorded in the handwritten sheets, but were incorrectly transcribed into the Excel sheet, eg Meter Number recorded as 041012**54**135 rather than 041012**45**135, or "C041-1012" was recorded as "<041-

1012".

- During the conditioning process, the database and the underlying network of Excel workbooks form a consistent and evolving 'tree' that represents the normalized data. Whilst undergoing conditioning, this 'tree' needs to be maintained, transported and manipulated as a whole.
- At some stage in the future, it is conceivable that the Access database could be 'disconnected' from the underlying network of Excel workbooks and the contained tables used in a 'standalone' way.

7. TABLE DESCRIPTIONS

7.1 TABLES - OVERVIEW

Table	Content	Comment
Aux_Dinokana_Lookup	Provides the information required to convert Dinokana Logger ID's based on a system derived from pole numbers (eg DN14A7), to a newer system of simplified ID's (eg 'DN04')	Simplified Logger ID's are required (for example) to facilitate upload to a GPS (6 character maximum)
Aux_Manuf_Data	Provides information about the various manufacturer codes used in STS meter numbers.	Not used at present
Aux_Matshana_Lookup_200206_200207	Provides the information required to convert Logger ID's for Matshana based on a system derived from pole numbers (eg M16DD11) used in the June 2002 Install survey, to a newer system of simplified ID's (eg 'MT01') used in the July 2002 Site survey.	
Aux_Matshana_Lookup_200206_2003	Provides the information required to convert Logger ID's for Matshana based on a system derived from pole numbers (eg M16DD11) used in the June 2002 Install survey, to a newer system of simplified ID's (eg 'MT01') used in the 2003 Site survey.	
Aux_GPS_Conv	Provides standardized Access functions to convert a variety of GPS source formats to a standard form of decimalized degrees	
Aux_Site_Details	Contains reference information (eg names, status, position etc) about various sites associated with Load Research activities, both past and present.	
Aux_Version_Log	A table containing a simple change log to track the development of the SR database.	
Aux_Work_Master	A manually-populated table containing details about the workflow associated with the SR database, eg which sites are included in the SR process, the survey dates, and the number of raw records collected during each survey.	Used to reconcile and check the SR process
Chanref_<SiteName>_<Year>_<Type>	A series of tables containing the dwelling-related survey information contained in the underlying Excel worksheets. The worksheets are dynamically linked to these tables.	Multiple tables Where the survey <Type> is not certain, it is omitted
Logref_<SiteName>_<Year>_<Type>	A series of tables containing the logger-related survey information contained in the underlying Excel worksheets. The worksheets are dynamically linked to these	Multiple tables Where the survey <Type> is not certain, it is omitted

Table	Content	Comment
	tables.	
Z_Out_Chanref_Allsites_Alldates	An output table containing all the multi-site, multi-year, processed survey references for all survey dates.	
Z_Out_Chanref_Allsites_Annual	As above, only where there are multiple references within a year, these have been coalesced into one single annual reference.	Under development - not complete

7.2 TABLES - DETAIL

For more detailed information concerning the tables above (eg field structure and formats), refer to the 'Design View' of the relevant tables in the Access database. Explanatory comments (such as those seen in the screenshots in section 5) can also be found there.

8. QUERY DESCRIPTIONS

8.1 QUERIES - OVERVIEW

In the descriptions that follow, the following conventions are used for naming :

C	Any single character
n	Any single numeral.
<Sitename>	One of 49 standardized site names
<Year>	A full 4-digit year, typically 2000 or later

Table	Content	Comment
A*cn* Present_Logref_<Sitename>_<Year>_<Type>	Presents the content of the associated Logger reference table for <Sitename> for <Year> in a standardized way.	eg A*cn* = AB1 <Type> = I (Installation) or S (Site). Where <Type> was uncertain, it is omitted
B*cn* Present_Chanref_<Sitename>_ <Year>_<Type>	Presents the content of the associated Channel (Dwelling) reference table for <Sitename> for <Year> in a standardized way.	eg B*cn* = BK1 <Type> = I (Installation) or S (Site). Where <Type> was uncertain, it is omitted
C*cn*_Union_<Sitename>_Chanref	Union queries used to join together references from successive Install and Site surveys within a site, into a single table of standardized format	
DA1 Union_<Sitename>_Chanref	Union query used to join together multi-site, multi-year Channel references into a single table of standardized format	These queries are written directly in SQL – name changes are not automatically updated
DB1 Union_<Sitename>_Logref	Union query used to join together multi-site, multi-year Logger references into a single table of standardized format	These queries are written directly in SQL – name changes are not automatically updated
DA2 Combined chanref data (dyna)	Presents the records returned by DA1 in a logical, history-based way, also adding additional key fields and data quality check indicators	
DA2a Chanref history for selected site (dyna)	As for DA2, but for a specific site only (desired site name entered as a parameter)	Invokes DA2. Site name needs to be as per standardized list
DB2 Combined Logref data (dyna)	Presents the records returned by DB1 in a logical, history-based way, also adding additional key fields and data quality check indicators	
DB2a Logref history for selected site (dyna)	As for DB2, but for a specific site only (desired site name entered as a	Invokes DB2. Site name needs to be as per standardized list

Table	Content	Comment
	parameter)	
EA1 Full refs (LID, Chan & MN)	Lists all references which are complete in the sense that they have non-null LoggerID, Channel and Meter Number.	
EA2 Full refs per key	Building on EA1, provides a count of all complete references against each unique key (eg AN01_1)	The count is incremented for each reference that takes place within a year (eg Install and Site), as well as for separate references in different years.
EA3 Count of full refs per site	Building on EA2, provides a count of all complete references against each site	
EB1 Count of linked references	Provides a count of the total number of reference records linked on a per-site basis	Used for reconciling work, ie expected vs actual record counts
FA1 All keys represented in complete references	Lists all keys (eg AN01_1) for which at least one complete reference exists, per site and per year	
FA2 Complete references by site and year	Provides a count of the number of complete references per site and per year	
FA3 Count of keys per site per year	Provides a count of the number of keys for which at least one complete reference exists, per site and per year	
FA4 Count of complete refs per represented key	Provides a count of the number of keys, and the number of complete references, per site and per year	Gives and indication of where multiple surveys took place in a specific site-year.
FA5 Count of keys for complete refs, by site and year	Provides a crosstab count of the number of complete references, per site and per year	
GA1 Average of dwelling positions, by site	Computes the average Latitude and Longitude of all references in a specific site as an indication of the approximate centroid	Does not take account of errors introduced by readings which may be erroneous
GA2 Offset of average positions from site centroids	With reference to the results of PB1, computes the offset (in metres) between the average position of all references in a site, and the centroid of the site as indicated in the table "Aux_Site_Details"	Used as an indicator of gross position errors
PB1Flush static site ref table	Deletes the content (not the structure) of the static table "Z Out Chanref_Allsites_Alldates" in preparation for rebuilding	
PB2 Make static site ref table	Constructs the static version of the complete site reference table ("Z Out Chanref_Allsites_Alldates") by appending to an existing structure	Adds fields containing related information (eg radius from associated site centroid and data quality check results)
BI0 through BI0d	Queries used to generate tables 'Aux_Matshana_Lookup_200206_200207' and 'Aux_Matshana_Lookup_200206_2003' for the purpose of mapping Logger ID's using different systems of reference.	
PC1 Combined site ref data (static)	Displays the content of the static site reference table in a logical way	
PC2 Incomplete references (C or MN missing)	Generates a list of references which have Logger ID's but one or both of Channel and Meter Number missing	
TA1 LP keys with non-null CN's	Lists the keys corresponding to those Logger Points that have a non-null Channel Number (and Logger ID)	
TA2 LP refs with non-null CN's	Lists all the reference data corresponding to the keys listed by TA1	
TA3 Latest MN's	For the references listed by TA3, selects the "latest, non-null" data by key	

Table	Content	Comment
TA4 Latest positions	For the references listed by TA3, selects the "latest, non-null" data by key	
TA5 Latest Eskom numbers	For the references listed by TA3, selects the "latest, non-null" data by key	
TA6 Latest names	For the references listed by TA3, selects the "latest, non-null" data by key	
TA7 Latest Erf	For the references listed by TA3, selects the "latest, non-null" data by key	
TA8 Latest non-null information - all sites		
TA9 Target list for socio visits	From the "latest, non-null"information generated by TA8, selects references for specific sites based on Key criteria	
TA9a Target list for socio visits - print version	As for TA9, but with positions output in DMDS	
TA9b Target list for socio visits - GPS version	As for TA9, but with positions output in DD and with Keys abbreviated to maximum 6 characters (only 2-character logger prefix) to suit GPS	
WD1 Worksheet recon	Presents a list of survey dates and raw record counts for those sites earmarked for inclusion in the SR database, along with reported survey dates and actual linked record counts, for the purpose of work reconciliation	
WD1a Work shortfalls	From the output of WD1, shows only those sites for which the reported record count is Null or less than the expected record count	

9. QUERIES – DETAIL

9.1 'A-series' queries – Presentation of logger records

The purpose of the A-series of queries is to present Logger reference data represented in the various linked tables in a standardized format, for later combination with reference data from other survey dates and other sites. The query naming convention is as follows :

A*cn*_Present_Logref_<SiteName>_<Year>_<SurveyType>

Where:

c	is a single character identifying the area
n	is a single digit identifying the year (typically one more than the year)
<SiteName>	is one of the 49 standardized site names
<Year>	is the year in which the survey was done
<SurveyType>	is a character indicating the type of Reference survey; where S=Site Reference and I=Installation Reference

Where the type of a survey was not certain (eg some of the early 2000 surveys), the <SurveyType> indicator is omitted.

These queries present or synthesize the following fields in order :

Field	Type	Description
Site	Text 50	Standardized site name

Field	Type	Description
Lat	Double 6	Latitude wrt WGS84 in DD format, ie ±dd.dddddd (South is negative)
Lon	Double 6	Longitude wrt WGS84 in DD format, ie ±dd.dddddd (West is negative)
PoleNo	Text 50	Number of ESKOM pole at which logger is located
LoggerID	Text 50	ID of logger
NumChans	Integer	Number of current channels for which logger is configured
SurvType	Text 50	Type or survey; one of Partial, Install, Site or Market
SurvDate	Date/Time	Date or survey dd/mm/yyyy
SheetNo	Text 50	Number of survey sheet on which logger's details were recorded
Comments	Text 255	Comments relevant to the logger entry

When compiling these queries, note that in order for the later survey- and site combination to be successful :

- The order of the fields as defined in the QBE grid must be the same for all queries, as given above
- The order of the fields as laid out in the presentation dynaset table must be the same for all queries, as given above
- The type of the underlying fields must be as given above

9.1.1 Notes pertaining to logger-presentation queries

- Where there is no data for a field (eg GPS positions in early surveys, the field must be set to the special Null value.
- For the Dinokana and Matshana sites, Logger ID's for some surveys have to be converted from a legacy (pole number-based) convention, to the simpler, modern naming convention. This is done by means of the lookup tables and 'Aux_Dinokana_Lookup ' and 'Aux_Matshana_Lookup_200206_200207' respectively, for the following years:
    - Dinokana : 2004 'Install' and 2005 'Site' surveys
    - Matshana : 2002 'Install' survey
- For surveys done in 2001, the single-character encoding of the 'TestType' field (A, B or C) was slightly different to that used in other years. For these 2001 surveys, the content of the 'TestType' field is converted to conform to a uniform coding scheme. The conversion scheme is documented in the table 'Aux_TestType_Coding' table. At this level (Presentation queries), a single character is still used. Later on (when generating static tables), a more meaningful coding scheme is used.

9.2 'B-series' queries – Presentation of channel records

The purpose of the B-series of queries is to present Channel reference data represented in the various linked tables in a standardized format, for later combination with reference data from other survey dates and other sites. The query naming convention is as follows :

B*cn*_Present_Chanref_<SiteName>_<Year>_<SurveyType>

Where:

c	is a single character identifying the area
n	is a single digit identifying the year (typically one more than the year)
<SiteName>	is one of the 49 standardized site names
<Year>	is the year in which the survey was done
<SurveyType>	is a character indicating the type of Reference survey; where S=Site Reference and I=Installation Reference

Where the type of a survey was not certain (eg some of the early 2000 surveys), the <SurveyType> indicator is omitted.

These queries present or synthesize the following fields in order :

Field	Type	Description
Site	Text 50	Standardized site name
LoggerID	Text 50	ID of logger
ChannelNo	Text 50	Channel number
Key	Text 50	<LID>_<C>_<Year>, eg
Bearing	Integer	Bearing of dwelling from pole at which logger is located
Lat	Double 6	Latitude wrt WGS84 in DD format, ie ±dd.dddddd (South is negative)
Lon	Double 6	Longitude wrt WGS84 in DD format, ie ±dd.dddddd (West is negative)
Erf	Text 50	Number of ESKOM pole at which logger is located
Eskom	Text 50	Number of current channels for which logger is configured
MeterNo	Text 50	
HolderName	Text 100	Surname, FirstName
SurvDate	Date/Time	Date or survey dd/mm/yyyy
Comments	Text 255	Comments relevant to the logger entry
SheetNo	Text 50	Number of survey sheet on which logger's detaisl were recorded
TestType	Text 3	A-, B- or D Test
SurvType	Text 50	Type or survey; one of Partial, Install, Site or Market

When compiling these queries, note that in order for the later survey- and site combination to be successful :

- The order of the fields as defined in the QBE grid must be the same for all queries, as given above
- The order of the fields as laid out in the presentation dynaset table must be the same for all queries, as given above
- The type of the underlying fields must be as given above

9.2.1 Notes pertaining to channel-presentation queries

- Where there is no data for a field (eg GPS positions in early surveys, the is set to the special Null value.
- For the Dinokana and Matshana sites, Logger ID's for some surveys have to be converted from a legacy (pole number-based) convention, to the simpler, modern naming convention. This is done by means of the lookup tables and 'Aux_Dinokana_Lookup ' and 'Aux_Matshana_Lookup_200206_200207' respectively, for the following years:
    - Dinokana : 2004 'Install' and 2005 'Site' surveys
    - Matshana : 2002 'Install' survey

9.3 'C*c*1-series' queries

The purpose of the C*c*1-series of queries is to combine the data presented by the underlying Channel Reference presentation queries (which arise from surveys done at the same site, but at different dates) into a single site with exactly the same format. For instance, the SQL source for the 'CA1 Union_Antioch_Chanref' query is as follows :

TABLE [BA1 Present_Chanref_Antioch_2000] UNION ALL

TABLE [BA2 Present_Chanref_Antioch_2001_S] UNION ALL
TABLE [BA3 Present_Chanref_Antioch_2002_S] UNION ALL TABLE [BA4 Present_Chanref_Antioch_2003_S];

This statement has the effect of combining the data for the surveys 2000, 2001 (Site), 2002 (Site), and 2003 (Site) into a single dynaset. The 'UNION ALL' keywords force all records to be used, ie no 'DISTINCT' logic is applied.

The result is a dynaset containing multiple records (corresponding to different surveys) for the same channel. Depending on the completeness of the underlying set, only some of the records will have all fields populated.

9.4 'C*c*2-series' queries

The purpose of the C*c*2-series of queries is to combine the data presented by the underlying Logger Reference presentation queries (which arise from surveys done at the same site, but at different dates) into a single site with exactly the same format. For instance, the SQL source for the 'CA2 Union_Antioch_Logref' query is as follows :

TABLE [AA1 Present_Logref_Antioch_2000] UNION ALL
TABLE [AA2 Present_Logref_Antioch_2001_S] UNION ALL
TABLE [AA3 Present_Logref_Antioch_2002_S] UNION ALL TABLE [AA4 Present_Logref_Antioch_2003_S];

This statement has the effect of combining the logger reference data for the surveys 2000, 2001 (Site), 2002 (Site), and 2003 (Site) into a single dynaset. The 'UNION ALL' keywords force all records to be used, ie no 'DISTINCT' logic is applied.

The result is a dynaset containing multiple records (corresponding to different surveys) for the same logger. Depending on the completeness of the underlying set, only some of the records will have all fields populated.

9.5 DA1 Union All Chanrefs

This query combines all of the Channel Reference records from the underlying 'C*c*1' queries (which are single-site, multiple-date records) into a single set of multiple-site, multiple-date records having exactly the same data format. The SQL source is as follows :

TABLE [CA1 Union_Antioch_Chanref] UNION ALL
TABLE [CB1 Union_Garagapola_Chanref] UNION ALL
TABLE [CC1 Union_Gasese_Chanref] UNION ALL
TABLE [CE1 Union_Maconqo_Chanref] UNION ALL
TABLE [CF1 Union_Mfazazane_Chanref] UNION ALL
TABLE [CG1 Union_Mafefe_Chanref] UNION ALL
TABLE [CI1 Union_Matshana_Chanref] UNION ALL
TABLE [CK1 Union_Tambo_Chanref] UNION ALL
TABLE [CL1 Union_Dinokana_Chanref] UNION ALL
TABLE [CM1 Union_Driekoppies_Chanref] UNION ALL
TABLE [CN1 Union_Khayalitsha_Chanref] UNION ALL
TABLE [CO1 Union_Vlaklaagte_Chanref] UNION ALL
TABLE [CP1 Union_Ikgomotseng_Chanref] UNION ALL

TABLE [CQ1 Union_Peacetown_Chanref];

9.6 DB1 Union All Logrefs

This query combines all of the Logger Reference records from the underlying 'C*c*2' queries (which are single-site, multiple-date records) into a single set of multiple-site, multiple-date records having exactly the same data format. The SQL source is as follows :

TABLE [CA2 Union_Antioch_Logref] UNION ALL
TABLE [CB2 Union_Garagapola_Logref] UNION ALL
TABLE [CC2 Union_Gasese_Logref] UNION ALL
TABLE [CE2 Union_Maconqo_Logref] UNION ALL
TABLE [CF2 Union_Mfazazane_Logref] UNION ALL
TABLE [CG2 Union_Mafefe_Logref] UNION ALL
TABLE [CI3 Union_Matshana_Logref] UNION ALL
TABLE [CK2 Union_Tambo_Logref] UNION ALL
TABLE [CL2 Union_Dinokana_Logref] UNION ALL
TABLE [CM2 Union_Driekoppies_Logref] UNION ALL
TABLE [CN2 Union_Khayalitsha_Logref] UNION ALL
TABLE [CO2 Union_Vlaklaagte_Logref] UNION ALL
TABLE [CP2 Union_Ikgomotseng_Logref] UNION ALL
TABLE [CQ2 Union_Peacetown_Logref];

9.7 Query 'DA2 Combined chanref data (dyna)'

Query DA2 simply presents the Channel Reference data generated by the underlying 'DA1' Union query in an intuitive fashion. Multiple references for the same Logger ID and Channel are presented in date order to facilitate easy year-on-year comparison.

9.8 Query 'DA2a Chanref history for selected site (dyna)'

Query DA2a is based on Query DA2, but restricts the dataset to a single site, which the user must specify. The site name must be drawn from the standardized list and spelled correctly for the query to return results properly. This query facilitates easy year-on-year comparison of channel records for the same site.

9.9 Query 'DB2 Combined logref data (dyna)'

Query DB2 simply presents the Logger Reference data generated by the underlying 'DB1' Union query in an intuitive fashion. Multiple references for the same Logger ID are presented in date order to facilitate easy year-on-year comparison.

9.10 Query 'DB2a Logref history for selected site (dyna)'

Query DB2a is based on Query DB2, but restricts the dataset to a single site, which the user must specify. The site name must be drawn from the standardized list and spelled correctly for the query to return results properly. This query facilitates easy year-on-year comparison of logger records for the same site.

9.11 'VB' query series for identifying field-specific Union errors

When dynasets arising from different surveys (sites and dates) are combined using Union queries,

problems may arise due to differences in field order and underlying field types. The error messages generated by Access as a result can be rather obscure. The best way to resolve these errors is to use a 'Long form' SQL query in which the underlying fields are explicitly combined. Records from several dynasets can then be combined one field at a time until the offending field(s) are identified. The 'VB' series of queries provide various examples which were used during development.

9.12 'E-series' informational queries

The 'E-series' are informational queries intended to provide summary information as to the number of loggers and channels represented in the database, as well as the number of references of each type, ie how many times each type of reference (Logger or Channel) has been surveyed.

EA1 Counts the number of Channel references by Site, Year and Survey type

EA2 Counts the number of Logger references by Site, Year and Survey type
EB2 Counts the number of Logger references by individual Logger ID

EB3 Counts the number of Loggers per site, and sums the number of references of those loggers
EB4 Presents the results of EB3 as a quotient indicating the approximate Logger reference frequencies on a per-site basis

EC1 Presents the number of Channel- and Logger references per Site, Year and Survey type
EC2 Sums the results of EC1 to give a per-site total

Note that all of these queries still work with dynasets which extend right back to the dynamically-linked Excel worksheets. The fully-expanded SQL statements become quite complex and computation-intensive. These queries represent the approximate point at which Access runs out of capacity – eg it is impossible to generate the results of EC2 in one step as the underlying SQL queries become 'too complex' to execute.

9.13 'FA' and 'FB' query series – gross positional errors

Queries FA1 and FA2 are designed to detect gross positional errors in reported positions of dwellings with respect to site centroids. Likewise, queries FB1 and FB2 perform the same function for loggers.

Query FA1 computes the average position of all Sites, by simple averaging the Latitudes and Longitudes of all dwelling positions as reported from the Channel references. Query FA2 then compares these average positions with a set of manually-specified site centroids as contained in the auxiliary 'Aux_Site_Details' table. The result is presented as the offset (in metres) between these two points. High offset values would indicate gross errors in the underlying data.

Queries FB1 and FB2 are structured identically, but operating on the logger data.

Note that these queries are very simplistic in that the average positions are themselves affected by any underlying errors in the data set. It is however useful in exposing simple transcription or GPS datum/conversion errors.

Note also that these queries operate on dynasets and are 'live', ie they immediately reflect changes in the underlying data and do not require any static table to be rebuilt.

9.14 Note : Breakpoint between queries based on dynasets and those based on static tables

The 'F' series of queries are the last that operate on dynamic data arising from linked spreadsheets. Queries which follow build, and then use, a static table containing the equivalent data. This is done to reduce processing overhead and increase speed. However, it does mean that the user must remember to rebuild the static table(s) after changes have been made to the underlying dataset.

9.15 Queries HA1 and 'HA2 Make static chanref table'

The table 'Z Out_Chanref_Allsites_Alldates' defines a data structure into which the amalgamated Channel references can be stored. Query HA1 flushes (ie deletes the contents of) this table. Query HA2 populates it, based on the Channel Reference records synthesized by the underlying linked spreadsheets and presentation queries. This allows subsequent, more complex queries to operate on a static table without the overhead of generating a new dynaset on every invocation.

For the most part, the fields generated by the associated query 'DA1 Union All Chanrefs' are simply carried into the corresponding fields in the static table 'Z Out_Chanref_Allsites_Alldates'. However, the following additional fields are synthesized at query run time :

Year — The 'SurvDate' (Survey Date) field is used to generate a 4-digit year field, which is useful on its own in later queries.

Frm_Cntrd — The centroid position of the site in which the dwelling resides is looked up in the table 'Aux_Site_Details' and used to compute the distance (in metres) between the reported dwelling position and the centroid itself.

Frm_Loggr — The position of the logger from which the dwelling is monitored is looked up in the virtual table (dynaset) generated by the query 'IC1 Average logger positions' and used to compute the distance (in metres) between the reported positions of the dwelling and the logger.

TestType — The character-based encoding scheme used to prepresent the type of test by which the logger connection was verified is converted to a more intuitive keyword-based encoding scheme through a lookup in the table 'Aux_TestType_Coding'. (Recall that the character-based scheme was normalized to a consistent convention in the Presentation queries).

STS_OK — For records having 11-digit meter numbers (ie STS meter numbers), a Visual Basic function called 'Luhn' computes the rightmost (11^{th}) digit that should be associated with the leftmost 10 digits. Where the actual digit agrees, this text field is set to the string 'Yes'. If the digit is incorrect, it is set to 'No'. For records having null or non-STS meter numbers, the field is set to the special 'Null' value.

Pos_OK — For records having GPS positions, this text field is set to 'Yes' if the dwelling position is within 8000m (8km) of the associated site centroid AND within 500m (0.5km) of the associated logger. Records with no GPS position have this field set to Null.

C_Key — This field is populated with a Logger- and Channel-specific text key of the form AN01_1.

CYM_Key — To the C_Key is added a 4-digit year and 2-digit month to form a Logger-, Channel-

and Month-specific text key.

9.16 Queries HB1 and 'HB2 Make static logref table'

The table 'Z Out_Logref_Allsites_Alldates' defines a data structure into which the amalgamated Logger references can be stored. Query HB1 flushes (ie deletes the contents of) this table. Query HB2 populates it, based on the Logger Channel Reference records synthesized by the underlying linked spreadsheets and presentation queries. This allows subsequent, more complex queries to operate on a static table without the overhead of generating a new dynaset on every invocation.

For the most part, the fields generated by the associated query 'DB1 Union All Logrefs' are simply carried into the corresponding fields in the static table 'Z Out_Logref_Allsites_Alldates'. However, the following additional fields are synthesized at query run time :

Year — The 'SurvDate' (Survey Date) field is used to generate a 4-digit year field, which is useful on its own in later queries.

Frm_Cntrd — The centroid position of the site in which the dwelling resides is looked up in the table 'Aux_Site_Details' and used to compute the distance (in metres) between the reported dwelling position and the centroid itself.

Pos_OK — For records having GPS positions, this text field is set to 'Yes' if the logger position is within 8000m (8km) of the associated site centroid. Records with no GPS position have this field set to Null.

CYM_Key — To the 'LoggerID' string (eg 'AN01') added a 4-digit year and 2-digit month to form a Logger- and Month-specific text key.

9.17 Queries IA1 and IB1 - 'Combined Chanref' and 'Combined LogRef' data (static)

These queries merely present the data in the associated static tables in an intuitive way, ie by Site, Logger, Channel (in the case of Channel references) and then by survey date. In this way, successive references of the same entity can be seen in chronological order along with the associated fields and error flags.

9.18 Query IC1 – Average logger positions

By operating on the static table 'Z Out_Logref_Allsites_Alldates', this query independently averages the latitude and longitude of every reference of each logger, to produce a time-averaged position for that logger.

The query is written to exclude any logger records for which the 'Pos_OK' flag is not set to "Yes", ie the averaged logger positions exclude any records where the position data is suspect. At the present time, this check is simplistic in that the Pos_OK flag is derived only from a check on whether the logger position is within 8km of the site centroid.

9.19 Query IC2 – Loggers with logrefs > 100m from average pos'n

By using the average logger positions computed by query IC1, this query selects logger records whose reported GPS position is over 100m from the time-averaged position for that logger. These

records indicate logger references that are potentially wrong in respect of reported positions.

Even though any wrong position reports for a logger themselves affect the computation of the time-averaged position for that logger (the point of reference), this simplistic query has nonetheless proven effective in identifying suspect logger references.

9.20 Query IC3 – Loggers > 100m from P_ave, with histories

Building upon the results of IC2, query IC3 presents those logger references identified as having GPS positions over 100m from their time-average positiond, together with other references for the same logger in chronological order. For additional reference, the query also presents the distance of the reported logger from the site centroid'. Large changes in one or both of these figures signify possible position reference errors which can then be traced back to source.

9.21 'ID' query series – Movement between logger position reports

Query ID1 computes the distance between a position reference for a given logger, and all other position references for the same logger. The minimum, average and maximum values of these 'movements' are presented per logger ID. Apart from giving a 'feel' for the 'repeatability'of GPS position references, large values for minima, maxima and averages point towards possible positon reference errors.

Building upon query ID1, ID2 allows the user to select an arbitrary 'maximum acceptable movement' between position references for the same logger, and then highlights those loggers which are associated with position references that exceed that limit. In examining the results presented by ID2, loggers with high 'Maximum-to-Minimum' or 'Maximum-to-Average' ratios probably indicate refernce groups with one suspect reading.

Note : 'Selective Availability' (deliberate degradation of the accuracy of the GPS system) was disengaged by the US Military on 2 May 2000. All GPS positions gathered as part of the Site Reference project would therefore not have been compromised by the effects of SA.

9.22 'JA' query series

The 'JA' series of queries (JA1 through JA9) provide summary information about the completeness of the channel references in the database in terms of 3 main fields only, namely :

- Logger ID,
- Channel Number, and
- Meter Number

By design, all records that form part of the channel reference dataset have a Logger ID. The number of reference records falling into each of the four possible states of the other two fields (Channel- and Meter Number), are enumerated by queries JA2 through JA5. The subtotals of these numbers, together with the total number of channel references and loggers, are summarized in different ways by queries JA6 through JA9.

As an example, the table below shows the content of a dynaset produced by query JA8 :

JA8 Chanref states and counts per site

Site	Num_LIDs	TotChanRefs	Tot_LID_C_MN	Tot_LID_C_xx	Tot_LID_x_MN	Tot_LID_x_xx
Antioch	32	371	222	119	5	25
Dinokana	19	243	181	26	6	30
Driekoppies	25	246	171	39	10	26

JA8 Chanref states and counts per site						
Site	Num_LIDs	TotChanRefs	Tot_LID_C_MN	Tot_LID_C_xx	Tot_LID_x_MN	Tot_LID_x_xx
Garagapola	34	290	193	90	2	5
Gasese	31	348	194	135		19
Ikgomotseng	24	190	141	35	11	3
Khayalitsha	14	267	182	32	8	45
Maconqo	22	129	97	15	4	13
Mafefe	33	174	10	160		4
Matshana	29	331	236	59	6	30
Mfazazane	29	220	158	38	11	13
Peacetown	24	173	126	42	1	4
Tambo	34	397	210	129	16	42
Vlaklaagte	20	270	209	20	20	21

As an illustration, In Antioch:

- 32 different logger ID's are represented in a total of 371 channel references
- Of these 371 channel references, all have Logger ID's, and
    - 222 have both Channel- and Meter Numbers,
    - 119 have Channel- but no Meter Numbers,
    - 5 have Meter- but no Channel Numbers, and
    - 25 have neither Channel- nor Meter Numbers

9.23 'JB' and 'JC' series of queries

Where the JA series generate summary data for Channel References, the JB series generate corresponding data for Loggers. The JC1 query combines the results of both Channel- and Logger-oriented data, an example of which is given below :

JC1 LID, Logref and Chanref counts per site				
Site	NumLoggers	TotLogRefs	TotPeakChans	TotChanRefs
Antioch	32	125	93	371
Dinokana	19	57	81	243
Driekoppies	25	75	82	246
Garagapola	34	124	79	290
Gasese	31	124	87	348
Ikgomotseng	24	47	92	190
Khayalitsha	14	42	89	267
Maconqo	22	43	65	129
Mafefe	33	64	87	174
Matshana	29	145	66	331
Mfazazane	29	87	73	220
Peacetown	24	48	87	173
Tambo	34	125	100	397
Vlaklaagte	20	60	90	270

As an illustration, In Antioch:

- 32 different logger ID's are represented in a total of 125 Logger references
- Based on history to date, 93 channels have been monitored by these 32 loggers, ie somewhat less than 3 channels per logger, on average. (Interestingly, the corresponding figure for Khayalitsha is 6.4 channels per logger)

- These 93 channels have been referenced a total of 371 times

Note that these figures correspond to those - derived somewhat differently - presented in query JA8.

9.24 'P' query series – Channel reference error tracing

All queries in the 'P' series are intended to highlight possible errors in the channel references for resolution at source.

Note that these figures correspond to those - derived somewhat differently - presented in query JA8.

9.24.1 PA1 and PA2 – Channel references with bad positions

By selecting those Reference records with the 'Pos_OK' flag not set to 'Yes' (PA1), and presenting them together with other references for the same Channel (PA2), it is usually possible to see which Reference record(s) of a chronological series are at fault. As at the time of writing, this query shows 48 'unfixable' records, as all other errors have been resolved at source.

9.24.2 PB1 and PB2 – Channel references with bad STS meter numbers

By selecting those Reference records with the 'STS_OK' flag not set to 'Yes' (PB1), and presenting them together with other references for the same Channel (PB2), it is usually possible to see which Reference record(s) of a chronological series are at fault. As at the time of writing, this query only shows 11 'unfixable' records, as all other errors have been resolved at source.

9.24.3 PC1 and PC2 – 'Strange' legacy meter numbers

Many legacy electricity meters (notably Conlog card-type meters) have non-STS meter numbers of the form Cnnnn-nnn. In the Reference worksheets, these were often captured with the 'C' replaced by an '<' and/or the hyphen omitted. These queries highlight such entries (PC1) and present them with other references for the same meter (PC2) so that the erroneous record(s) can be found and corrected.

9.24.4 PD1 and PC2 – Dwellings 'far' from site centroid

Query PD1 uses the previously-generated 'Frm_Cntrd' distance to identify dwellings with reported positions which are 'far' (over 2km) from the centroid of the associated site.

Query PD2 presents those records identified by PD1, along with other references for the same channel, in chronological order. Large changes in the 'Frm_Cntrd' records with generate Minimum, Maximum, Average, Standard Deviation (SD) and 'Scatter' (SD/Ave) figures for all dwelling positions by site.

refere and By selecting those Reference records with the 'STS_OK' flag not set to 'Yes' (PB1), and presenting them together with other references for the same Channel (PB2), it is usually possible to see which Reference record(s) of a chronological series are at fault. As at the time of writing, this query only shows 11 'unfixable' records, as all other errors have been resolved at source.

9.24.5 PD2 – 'Scatter of dwellings from site centroid'

Query PD1 uses the previously-generated 'Frm_Lggr' distance to generate Minimum, Maximum, Average, Standard Deviation (SD) and 'Scatter' (SD/Ave) figures for all dwelling positions by site.

refere and By selecting those Reference records with the 'STS_OK' flag not set to 'Yes' (PB1), and presenting them together with other references for the same Channel (PB2), it is usually possible to see which Reference record(s) of a chronological series are at fault. As at the time of writing, this query only shows 11 'unfixable' records, as all other errors have been resolved at source.

9.25 'PG' query series – duplicate channel entries

These queries identify (PG1) and present with other records for the same channel (PG2), Reference records having the same 'CYM_Key' (LoggerID, Channel, Year and Month) as one another. Such records usually arise from duplicate rows in the underlying Excel spreadsheets.

As at the time of writing, this query generates no records, as all such errors have been resolved at source.

9.26 'SA' query series – coalescence of multiple intra-year records

- SA1 selects only channel reference records that have a valid channel number
- SA2 selects those that have multiple references in the same year
- SA3 selects the records (2 or more) that correspond to these channel keys
- SA4 finds the earliest and latest dates of the records concerned
- SA5 extracts the data fields corresponding to the earliest date
- SA6 extracts the data fields corresponding to the latest date
- SA7 selects from these the 'most likely' data for each field for that year, using a VB function called 'Nullifier'

The 'Nullifier' function uses the following logic :

Earliest value	Latest value	Result	Comment
Null	Null	Null	No information
Null	Not Null	Latest Value	Believe the latest data
Not Null	Null	Earliest Value	Believe the only data
Equal		Latest Value	No problem
Not Equal		Null	Conflict – best assume nothing

Query SA8 takes the arbitrated values for the 'main' fields and appends them to a table with pre-defined structure. At the time of writing, the outputs are not utilized – further development is needed here.

9.27 'TA' query series – generating dwelling lists for socio-demographic surveys

The TA queries are designed to extract channel (dwelling) reference data in a form suitable for generating lists to direct researchers doing socio-demographic surveys. The main points to bear in mind here are :

- Commercial GPS units generally only accept 6-character waypoint names

- The SD surveyors prefer dwelling coordinates to be in DDM format

The sub-queries operate as follows:
- TA2 selects the latest Metere multiple references in the same year
- TA3 through TA7 select the most recent, non-null data (5 main fields)
- TA8 groups these selected fields
- TA9 selects only records from certain sites (typically active ones)
- TA9a converts the TA9 list into a suitable print form, eg presents positions in DDM format
- TA9b converts the TA9 list into a GPS-ready form, eg 6-character Logger ID's as waypoint names

9.28 'VA' query series for Matshana pole-matching

Between the 2002 Install and Site surveys for Matshana, a different system of Logger ID numbering was introduced, with formats as follows :

2002 Install survey	Of the form "M07DE04" (presumably based on pole numbering)
2002 Site survey	Of the form "MT01"

To enable site reference reporting on a common basis, the Logger ID's reported in the 2002 Install survey had to be mapped to those in the later Site survey. Unfortunately, the original cross-reference (lookup table) was lost, hence the "AI" set of queries finds matches based on proximity of the reported logger positions.

9.28.1 VA1 Find Matshana inter-logger distances

This query calculates the distance between every pole in the Site survey with respect to every pole in the earlier Install survey. The distances are presented (per Site-surveyed pole number) in descending order. The lowest distance therefore suggests the most likely match to a pole in the earlier Install survey.

9.28.2 VA2 Find lowest Matshana inter-logger distances

Building on the distances calculated in the previous query, this query selects the Install pole that is closest to each Site pole.

9.28.3 VA3 Match Matshana loggers

Building on the results of the previous query, this query matches the Logger ID's from the 2002 Install and Site reference surveys, and displays the distance between the reported logger (pole) positions.

By changing it from a 'Select' to a 'Make-Table' query, AI0b was used to construct the table 'Aux_Matshana_Lookup_200206_200207', which is a static table allowing a user to look up the new-format Logger ID (of the form MTxx) corresponding to an older 'pole-style' Logger ID.

9.28.4 VA4 Duplicate Matshana logger matches (02-02)

Essentially an error-checking query, this query counts the number of 'Install' (2002_06) loggers listed as matching a 'Site' (2002_07) logger. A count greater than 1 indicates that a new 'Site' Logger ID has been matched (based on lowest distance) to more than 1 'Install' LoggerID.

Any duplicate matches in the 'Aux_Matshana_Lookup_200206_200207' lookup table have already been resolved by direct editing.

9.28.5 VA5 Duplicate Matshana logger matches (02-03)

Essentially an error-checking query, this query counts the number of 2003 'Site' loggers listed as matching a 'Site' (2002_07) logger. A count greater than 1 indicates that a new 'Site' Logger ID has been matched (based on lowest distance) to more than 1 'Install' LoggerID.

Any duplicate entries have already been resolved.

9.29 CHANNEL REFERENCE PRESENTATION QUERIES

This 'B' series of queries are designed to "present" the data collected from site surveys in a standardized way, for later collation into a single table with uniform structure.

The data contained in each of the site reference surveys is conditioned to include fields and formats as specified in the table below :

Field name	Description
Site	The name of the site where the reference was taken
LoggerID	The ID of the logger at the LP
ChannelNo	The channel number of the logger at the LP
Key	A unique key identifying the Logger ID and Channel in the form PP(P)LL_CC
Bearing	Bearing of the dwelling with respect to the pole at which the logger is located
Lat	GPS Latitude of the LP-monitored dwelling, wrt WGS84 and in the format dd.ddddd
Lon	GPS Longitude of the LP-monitored dwelling, wrt WGS84 and in the format dd.ddddd
Erf	Erf number of the ground on which the LP-monitored dwelling is situated
Eskom	ESKOM number of the LP-monitored dwelling
MeterNo	Meter Number of the LP-monitored dwelling
HolderName	The name (either house head or respondent) of the dwelling occupant
SurvDate	The date on which this reference was taken
Comments	Any relevant comments. Comments added during data purification are generally followed by the text '[JC]'
SheetNo	Number of the sheet within this site reference on which this LP-monitored dwelling's details are recorded
TestType	Character indicating type of test : A=Load at house (surest), B=Load at pole (only in 2001 - rare); C=wiring

Field name	Description
	trace (least certain)
DKey	A unique key identifying the Logger ID, Channel and Year/Month of this reference
Year	Year in which the reference was taken
MN_Tag	Meter Number check tag - 'Good', 'Bad' or Null depending on whether STS meter number passes Luhn check
Pos_Tag	Position check tag - 'Good', 'Bad' or Null depending on whether position lies sensibly close to site centroid
C_Lat	Latitude of the centroid of the site in which this dwelling is situated
C_Lon	Longitude of the centroid of the site in which this dwelling is situated
Radius	Distance of dwelling from area centroid in metres
SurvType	Type of survey, ie Installation Refererence, Site Reference or Market Research

Note the following :

Lat/Lon — Records for which no GPS position was available (eg early 2000 surveys) have these fields set to a 'Null' value. Records in non-standard formats (eg Degrees / Minutes / Decimal Seconds (dd mm ss.s) or Degrees / Decimal Minutes (dd mm.mmm)) are converted to decimal degrees.

LoggerID — Only records with non-null LoggerID's are accepted, the assumption being that a record without this basic datum is essentially useless.

ChannelNo — Records with no channel number are allowed. In this case, this field is set to a 'Null' value.

HolderName — The Respondent or House Head name is standardized to the form 'Surname, Name' or 'Surname, <Initial>

The 'B' series of queries are named according to the following convention <Prefix>_Present_Chanref_<Sitename>_<Year>_<Type>, where :

Prefix — A query prefix including the unique table/query prefix allocated to that site. Loosely, the last digit is one more than the survey year.
Chanref — Indicates that the dynaset table generated relates to channel references.
Sitename — The standardized site name.
Year — The year in which the survey was conducted.
Type — Where necessary (viz multiple audits within a year), the letters "I" or "S" are used to indicate either an Installation or a Site reference

Example: The query name "AI3 Present Chanref Matshana 2002_S" indicates a query used for conditioning of channel reference data derived from a Site reference for the Matshana site in 2002.

10. APPENDIX : 'RECIPE' FOR ADDITION OF SITE REFERENCE DATA

10.1 BASIC PROCESS

The process given below can be applied for both Logger and Channel reference data ('Master' and 'Detail' worksheets respectively). The procedure is described for Loggers, but directives for Channel imports are given in parentheses.

- Examine the structure of the 'Master ('Detail') worksheet in the relevant site reference workbook. If the structure is unsuitable for import into Access (eg multi-row header), create a copy - typically 'Master(2)' ('Detail(2)') and make the required changes.
- Check that the format of data (eg GPS positions) in the same fields (cells) is uniform across all records. If not, normalize.
- Check that the cell data type is the same within columns, eg all ESKOM numbers stored as numbers or as text. Normalize to one type. (Note : this step can be done later)
- Link the spreadsheet to the Access SR DB using Access's 'Get External Data' feature. Once imported, name the table appropriately.
- Once imported, check the resulting table. Look for 'Num#' error indicators, blank rows etc and resolve at source.
- From the selection of Presentation query templates, select an appropriate template and edit to suit the field names and formats in the newly-linked table. Be careful to keep the field order in the QBE grid and the resulting dynaset table the same, as this affects the success of subsequent shortform Union queries.
- Name the presentation query appropriately and run it to check results . Refine as required.
- In the Union query for the associated site, append to the SQL text (using the UNION ALL command) to include the presentation query for the newly-linked site. Run the Union query and check the results. Problems caused by inter-year format differences will be evident and should be resolved here.
- Run the 'DB1' ('DA1') union query and check that the augmented site combines properly with other sites. An increase in the total record count should be visible.
- Flush the relevant static table by running query 'HB1' ('HA1'), and then re-generate by running 'HB2' ('HA2').
- Run analysis queries and resolve problems at source by editing the underlying Excel worksheets.

10.2 TIPS FOR PROCESSING

- Blank rows : Linking Excel worksheets into Access tends to result in tables with several trailing blank rows. This can be cured by deleting the required number of rows following the non-blank rows in the Excel sheet.
- Mixed number/text record fields : In some years, the content of specific fields (eg ESKOM number) in all records may be purely numeric (eg 2230, 123); whilst in other years there is a mixture of numeric and text data (eg 2230, 2251B). Generally, Access will determine the type of the table field at link time, ie each time the table is accessed. When Access types a field as 'Numeric' and then encounters a non-numeric value, a '#Num' error value is displayed in the associate table. To cure this, one must ensure that the type of data contained in a specific field is uniform across all records, eg all numeric or all text. This also prevents errors when multiple tables are combined. As a rule, it is best to standardize on the Text type for all such fields.

10.3 GENERAL TIPS

- Access 2003's "Name AutoCorrect" facility tracks changes made to queries and tables and updates affected entities automatically. However, it does not operate for queries written directly in SQL, eg the "C" and "D" series of Union queries.