Paper 020-30

Rule Your Data with The Link King©
(a SAS/AF® application for record linkage and unduplication)

Kevin M Campbell, DrPH
Washington State Division of Alcohol and Substance Abuse

ABSTRACT

Administrative datasets containing client identifying information (names, birthdates, SSNs) are often used for a variety of research and evaluation projects. The projects often require the linking of two or more independently maintained client rosters in order to track service utilization across different systems. Unfortunately, a given client may be represented with slightly different identifying information both within and across administrative datasets. Discrepancies arise from a variety of reasons including:

- Use of nicknames
- Hyphenated names
- Misspelled names
- Transposed SSN digits
- Transposed date fields

Failure to identify and appropriately deal with this problem may lead to incomplete linking of client records and, ultimately, introduce unnecessary error into the research or evaluation project.

This paper introduces The Link King - a SAS/AF application for use in the linkage and unduplication of administrative datasets. The Link King features a data importing and formatting wizard, artificial intelligence to insure appropriate linking protocols are used, a powerful interface for manual review of "uncertain" linkages, an ability to generate random samples of links for validation, and easy "point-and-click" editing of the final roster of consolidated records.

Visit www.the-link-king.com for more information about this public domain software or to download The Link King.

RECORD LINKAGE AND CONSOLIDATION ALGORITHMS

There are two approaches to the linkage and unduplication of client identifiers in administrative datasets: deterministic linking and probabilistic linking.

Probabilistic linking is accomplished through the application of sophisticated statistical analysis. Ultimately, a formula is derived which generates a score for each record pair and cut points to identify “definite” matches, “possible” matches, and “non matches”. The formula incorporates weights specific to each of the data elements and scaling factors for many of the data elements. The weights reflect the relative importance of specific data elements in predicting a match. The scaling factors adjust the weights for a given record pair based on the “rarity” of the data value. For example, the scaling factor for the last name “Freud” would be much larger than that for the last name “Smith”.

The probabilistic algorithms used by The Link King were developed by MEDSTAT for the Substance Abuse and Mental Health Administration’s (SAMHSA) Integrated database project.[1]

Deterministic linking is accomplished by establishing specific criteria about what combination of data elements need to “match” and quality of the “match” in order to accept the link as valid. For example, one criterion to consider two client records a “match” might be that all of the following conditions must be met:

First Names: Must have an Approximate String Match Algorithm score of .75 or Higher

Last Names: Must have an Approximate String Match Algorithm score of .75 or Higher

[1] The Technical Monograph and original SAS program code are available for download at www.the-link-king.com

Middle Initial: Must be an exact match or be missing

SSN: Must have at least 7 digits with exact positional match

Birth date: Must be an exact match

Deterministic record linkage is often portrayed as a method which doesn't account for missing values and partial agreements and yields less success than probabilistic methods. For example, Whalen et. al.[2] believe that "probabilistic matching produces more links than other methods and that many of these links are missed by other methods. This indicates probabilistic linking routines are more accurate than other routines for matching person-level data."

This is not necessarily true. An intricate deterministic algorithm can be as successful – or more successful – than probabilistic algorithms in identifying valid links. The Link King's deterministic algorithms take into consideration partial matches for names, birthdates, and social security numbers as well as the "rarity" of names being compared and, depending on the extent of similarity across data elements, links records at one of 4 levels of certainty. The deterministic algorithms used in The Link King were developed at Washington State's Division of Alcohol and Substance Abuse for use in a variety of program evaluation and research projects.[3]

The most powerful tool for record linkage and unduplication is one that incorporates both deterministic and probabilistic algorithms as The Link King does.

MANAGING UNCERTAINTY IN RECORD LINKAGE AND CONSOLIDATION

Deterministic and probabilistic algorithms classify linkages along a continuum. At one end are "definite" matches. At the other end are "definite" non-matches. The remaining linkages contain discrepancies that lead to varying degrees of uncertainty about the appropriateness of the linkage. The Link King provides the user with a variety of tools to "manage" the uncertainty inherent in this process.

GENDER IMPUTATION

At a minimum, The Link King requires first name, last name and either date of birth or social security number. The accuracy of record linkage and consolidation improve as more data elements are included (e.g., middle name or initial, maiden name, and gender). To maximize the amount of information available during record linkage, The Link King features "gender imputation". If the input dataset is missing data for gender for any of the records, The Link King will – whenever possible – determine the gender of the individual associated with the record in question and fill in the missing data. Using death records from Washington State for 1980 through 2003, we have created a look-up-table that maps first names to gender. This look-up-table is used to determine the gender of the individual with missing data. To avoid incorrect gender imputation for ambiguous names like "Chris", "Pat", etc., the look-up-table only contains first names where only one gender was found in death records for that first name.

INVALID VALUE IDENTIFICATION

In some administrative datasets default values are assigned to cases with missing SSNs, birthdates, and names. If these defaults values are not identified and eliminated, they can generate bad links. For example, if a value of '999999999' was entered whenever the SSN was unknown, this value must be identified as an invalid value to prevent the program from incorrectly finding a "match" for SSN when 2 records contain this value.

In addition to automatically identifying impossible values for SSNs, The Link King generates a listing of birth dates and SSNs that occur most frequently (the top 2%) in input datasets. The user can quickly browse this listing and flag invalid values for birth dates and SSNs. The Link King will then update the input dataset, recoding these flagged values "missing".

SMART USE OF "SIMILARITY" IN MATCHING NAMES

Most record linkage and unduplication software (including The Link King) use "phonetic equivalence" or "spelling distance" as a means to identify misspelled names. The Link King uses an Approximate String Matching algorithm, SAS's "spedis" function, SAS's "soundex" function, and New York State Intelligence Information System's phonetic equivalence function.

Use of spelling distance algorithms and phonetic similarity can be very useful in identifying pairs of names that are - for all practical purposes - identical but these methods can sometimes lead to false matches. For example, Michael

[2] Whalen D., Pepitone A., Graver L., Busch J., "Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies", U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES, Substance Abuse and Mental Health Services Administration, July 2001

[3] Specific details about the deterministic criteria can be found in The Link King user manual, available for download at www.the-link-king.com

and Michelle are phonetic equivalents.[4] The Link King provides an interface through which the user can "teach" The Link King to treat name pairs found to be “similar” in a specific manner. Essentially, this is a mechanism for fine tuning “phonetic equivalence” or “spelling distance” algorithms for the current and future unduplication jobs.

NICKNAME IDENTIFICATION

In addition to “phonetic equivalence” or “spelling distance”, The Link King determines if the names being compared are nicknames. The Link King will treat names like “Bill”, “William”, “Billy”, “Will” etc. as virtual equivalents. A look-up-table of nicknames is provided with The Link King as well as an interface for easy addition of new nicknames to the table.

EASY MODIFICATION OF MINIMUM PROBABILISTIC WEIGHT VALUES

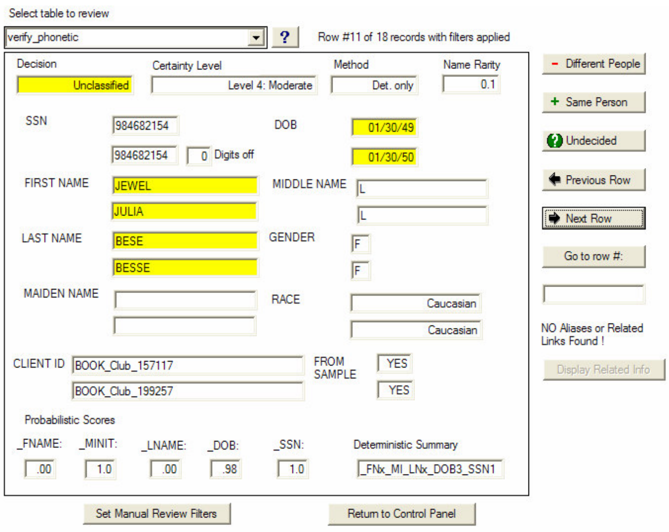
MEDSTAT’s probabilistic protocol utilizes two weights for each data element used in linking an “agreement” weight and “disagreement” weight. MEDSTAT has incorporated default values for minimum weights. The Link King empowers the user to (a) quickly and easily change the default minimum weights or (b) allow the program to use empirically determined weights, discarding the use of any mandated minimum value.

POWERFUL INTERFACE FOR MANUAL REVIEW UNCERTAIN LINKS

It is important that record linkage and unduplication software provide an interface for focused manual review of these uncertain linkages. Focused manual review refers to a review interface that allows the user to focus attention on links with similar types of discrepant information and to change the focus of the review quickly and easily. For example, one might initially want to focus the review on cases where the match was based - to some extent – on the phonetic equivalence of names. Later, the user might want to focus on uncertain cases where the link was based on similarity according to the Approximate String Matching Algorithm. Focused review is aided by an interface that highlights discrepant fields for easy identification.

The Link King provides the user with a powerful interface for easy review and classification of “uncertain” linkages. Discrepant fields are highlighted in yellow to aid in the review process.

Figure 1



Select table to review: verify_phonetic

Row #11 of 18 records with filters applied

Decision: Unclassified
Certainty Level: Level 4: Moderate
Method: Det. only
Name Rarity: 0.1

SSN: 984682154
984682154 | 0 Digits off
DOB: 01/30/49
01/30/50

FIRST NAME: JEWEL
JULIA
MIDDLE NAME: L
L

LAST NAME: BESE
BESSE
GENDER: F
F

MAIDEN NAME:
RACE: Caucasian
Caucasian

CLIENT ID: BOOK_Club_157117
BOOK_Club_199257
FROM SAMPLE: YES
YES

Probabilistic Scores

_FNAME:	_MINIT:	_LNAME:	_DOB:	_SSN:	Deterministic Summary
.00	1.0	.00	.98	1.0	_FNx_MI_LNx_DOB3_SSN1

NO Aliases or Related Links Found !

RANDOM GENERATION OF LINKAGE SAMPLES FOR VALIDATION

Manual review of all uncertain linkages is often not possible in extremely large linkage and unduplication jobs. In such a case, it is important to provide the user with information about the extent of uncertainty in linked records. Further, it is important to provide the user with a mechanism (based on empirical data) for selecting only those links where the degree of uncertainty is acceptable to the user. The Link King classifies linked records into one of 11 categories based on a) the protocol that established the link (deterministic, probabilistic, or both) and b) the degree of uncertainty in the linked records. The user can generate random samples of linked records in each of these 11 categories and, based upon the degree of error found in the random samples, choose to include (or exclude) any of the 11 categories.

When uncertain pairs are manually reviewed, random samples of validated links can be generated by the data manager to spot check decisions made by staff assigned to conduct the manual review.

EASY EDITS TO THE FINAL LINKAGE MAP

The linkage “map” consists of the consolidation of records believed to represent the same person under a common “uniqueid”. This group of consolidated records includes records that were directly linked to each other and, in some cases, records that were indirectly linked together. The process of gathering indirect links into a consolidation is called “chaining” and is illustrated below where records #1 and #3 have been “chained” together.

[4] It is important to remember that a “match” on a given name element (e.g., first name) is only one of many criteia in the decision to link records.

Record #	First Name	Middle Initial	Last Name	Maiden	Birthdate	SSN
1	KARAN	L	LONG		10/06/57	980056365
2	KAREN	L	LONG	SMITH	10/26/57	980056365
3	KARIN		SMITH		10/26/57	

Records #1 and #2 are directly linked due to compelling similarity in all fields.

Records #2 and #3 are directly linked due to compelling similarity on all available data fields.

Records #1 and #3 would not have been linked together if not for their common relationship with record #2. Therefore, records #1 and #3 have been "chained" or "cross-linked" together.

The Link King selectively chains records in his construction of the final linkage map. In the vast majority of cases the consolidation will meet with the user's approval. The Link King allows for the possibility of disagreement with the user and empowers the user to modify any group of consolidated records with a few mouse clicks. In the above example, the user could easily disaggregate record #3 from the grouping.

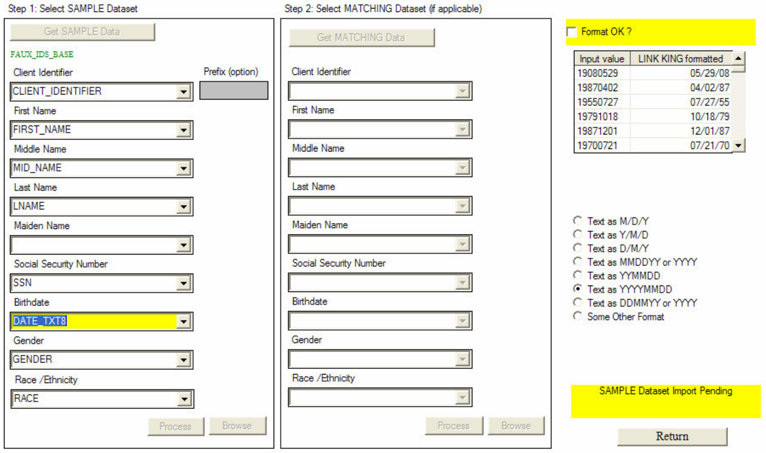
LINKING FOR KNAVES: MAKING IT EASY

Users of The Link King need no SAS programming experience. All functions are fully automated, including the "launching" of the program with a desktop icon.

PRESTO! DATA IMPORTING AND FORMATTING HAS NEVER BEEN EASIER

Data processing courtesy of "Presto" the data importing and formatting wizard virtually eliminates the need for any formatting of your input dataset. There are no naming conventions for variables and no special formatting requirements for gender, race, or birth date. Birthdates can be formatted as SAS dates, datetimes, or a variety of text string formats. Simply navigate to the location of the input dataset(s) and …Presto ! It's like magic.

Figure2



Step 1: Select SAMPLE Dataset

FAUX_IDS_BASE

Client Identifier	CLIENT_IDENTIFIER
Prefix (option)	
First Name	FIRST_NAME
Middle Name	MID_NAME
Last Name	LNAME
Maiden Name	
Social Security Number	SSN
Birthdate	DATE_TXT8
Gender	GENDER
Race /Ethnicity	RACE

Step 2: Select MATCHING Dataset (if applicable)

Client Identifier	
First Name	
Middle Name	
Last Name	
Maiden Name	
Social Security Number	
Birthdate	
Gender	
Race /Ethnicity	

□ Format OK ?

Input value	LINK KING formatted
19080529	05/29/08
19870402	04/02/87
19550727	07/27/55
19791018	10/18/79
19871201	12/01/87
19700721	07/21/70

○ Text as M/D/Y
○ Text as Y/M/D
○ Text as D/M/Y
○ Text as MMDDYY or YYYY
○ Text as YYMMDD
● Text as YYYYMMDD
○ Text as DDMMYY or YYYY
○ Some Other Format

SAMPLE Dataset Import Pending

ARTIFICIAL INTELLIGENCE INSURES APPROPRIATE PROCESSING OF DATA

Record linkage and consolidation tasks take a variety of forms. There is no single process that is appropriate for all jobs. Consider the following 2 scenarios:

Example A: Washington State's Division of Alcohol and Substance Abuse (DASA) needs to determine how many of their clients received services from the Mental Health Division (MHD). Both DASA's and MHD's client roster may contain multiple records for a given client with subtle variations on the client's legal name.

Example B: Washington State's Medical Assistance Administration (MAA) needs to determine how many of their client have been arrested by the Washington State Patrol (WSP). MAA's stringent requirements for verification of eligibility virtually guarantees there is no duplication of client's in their client roster. WSP's client identifier is fingerprint based but the identifying information contains many fictitious aliases (i.e., identifying information given in a deliberate attempt to deceive the person collecting the data).

A full discussion of the appropriate processes for each of these jobs cannot be presented here. In sum, Example A requires the unduplication of both datasets during the linking process. Example B, requires that neither dataset be unduplicated prior to linking. Other situations may require that only one of the datasets be unduplicated during the linking process.

The Link King has been endowed with artificial intelligence that virtually insures the job is processed in the most efficient and appropriate manner. Occasionally, The Link King may ask the user one or two "yes/no" questions to refine his recommendations.

BATCH PROCESSING

The time involved in record linkage and consolidation depends on a variety of factors including processor speed and size of datasets being linked/consolidated. Large unduplication jobs may take several hours to complete. The Link King features both interactive and batch processing modes. Once the user is familiar and comfortable with The Link King's processing, resource intensive jobs can be run over lunch or over night.

ABUNDANT OPTIONS

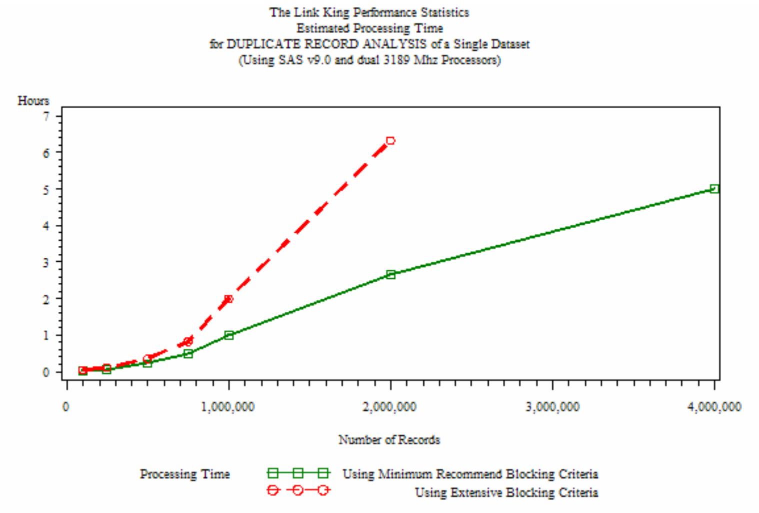
The Link King offers the user a number of options for customizing the linkage experience. These options allow the user to control required resources and specify acceptable levels of linkage uncertainty. Users need not take advantage of these options (or be intimidated by them). A perfectly adequate (Dare I say exceptional ?) linkage solution can be obtained through the use of The Link King's default settings. The default settings provide the user with a "conservative" solution, consolidating only those records where the certainty of the linkage is very high.

PERFORMANCE STATISTICS

Performance statistics vary depending on processor speed, degree of duplication in dataset being processed, and blocking criteria selected. The graph to the right illustrates processing time required for unduplication of a single dataset ranging in size from 10,000 to 4,000,000 records. The datasets used in this example contained approximately 25% duplicate clients. As expected, processing time was significantly higher when more extensive blocking criteria was used. For most tasks, the minimum recommended blocking criteria is adequate as it captures 95-97% of links.

Hard disk requirements for these tasks range from 20 gigs for unduplication of 1,000,000 records to 150 gigs for unduplication of 4,000,000 records.

Figure 3



GRAND TOUR: THE KINGDOM OF LINKS

Space limitations do not allow full exploration of the majesty that is The Link King although a detailed user manual is available for download. The following example of a "typical" linkage job using The Link King's default settings with no manual review illustrates the most basic operation of The Link King. This streamlined application is accomplished in 3 easy steps.

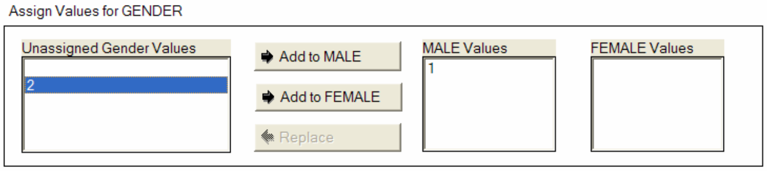
STEP 1: IMPORT AND FORMAT DATA

After launching The Link King, the user selects "Get New Data" from the main menu to access the data importing and formatting wizard (figure 2). Using the "Get Sample Data" or "Get Matching Data" controls, the user navigates to the location of the SAS dataset containing the identifier files to be used in the linkage. Variables to be used in the linkage process are identified through the use of convenient drop-down list boxes .

If categorical variables such as gender and race/ethnicity are to be used in the linking process, "Presto" - The Link King's data importing and formatting wizard – will guide the user through the application of comparable formats to these data elements. For example, if gender is coded a "0" for male and "1" for female in one dataset but "1" for male and "2" for female in the other dataset, a comparable format must be imposed on the two datasets.

The user can quickly and easily assign gender values to specified categories using the controls displayed in Figure4. A similar control is available to assign race/ethnicity values to specific categories. The format structure can be saved for use in other linkage jobs.

Figure 4


Assign Values for GENDER

Unassigned Gender Values	MALE Values	FEMALE Values
2	1	

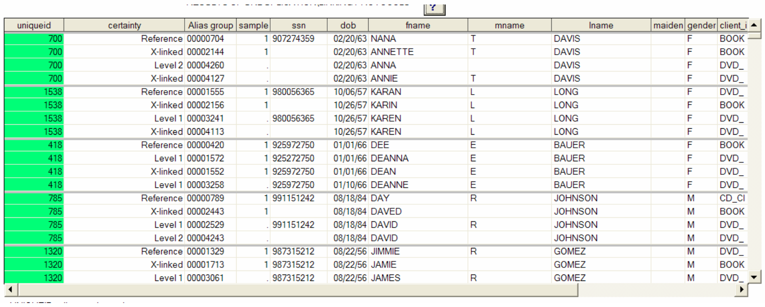
STEP 2: IMPLEMENT LINKAGE PROTOCOLS

After the data has been imported, the user navigates to the "Record Linkage and Unduplication Control Panel and – using the Batch Processing controls - selects "Run Entire Process". Interactive processing is also available which allows for considerable user refinement of the linkage process including modification of minimum values for probabilistic weights, manual review of "uncertain" linkages, and exclusion of linkages based on probabilistic scores, linkage method (deterministic or probabilistic), and "certainty" of the linkage decision. See The Link King's user manual for full details.

STEP 3: REVIEW/EDIT FINAL LINKAGE MAP

The user may view the final map of linkages - where all records determined to represent a single individual are consolidated under a common unique identification number - in a variety of sorted configurations. If the user believes that records have been incorrectly linked, edits are easily made with a few mouse clicks.

Figure 5



uniqueid	certainty	Alias group	sample	ssn	dob	fname	mname	lname	maiden	gender	client_i
700	Reference	00000704	1	907274359	02/20/63	NANA	T	DAVIS		F	BOOK
700	X-linked	00002144	1		02/20/63	ANNETTE	T	DAVIS		F	BOOK
700	Level 2	00004260	.		02/20/63	ANNA		DAVIS		F	DVD_
700	X-linked	00004127	.		02/20/63	ANNIE	T	DAVIS		F	DVD_
1538	Reference	00001555	1	980056365	10/06/57	KARAN	L	LONG		F	DVD_
1538	X-linked	00002156	1		10/26/57	KARIN	L	LONG		F	BOOK
1538	Level 1	00003241	.	980056365	10/26/57	KAREN	L	LONG		F	DVD_
1538	X-linked	00004113	.		10/26/57	KAREN	L	LONG		F	DVD_
418	Reference	00000420	1	925972750	01/01/66	DEE	E	BAUER		F	BOOK
418	Level 1	00001572	1	925272750	01/01/66	DEANNA	E	BAUER		F	DVD_
418	X-linked	00001552	1	925972750	01/01/66	DEAN	E	BAUER		F	DVD_
418	Level 1	00003258	.	925972750	01/10/66	DEANNE	E	BAUER		F	DVD_
785	Reference	00000789	1	991151242	08/18/84	DAY	R	JOHNSON		M	CD_CI
785	X-linked	00002443	1		08/18/84	DAVED		JOHNSON		M	BOOK
785	Level 1	00002529	.	991151242	08/18/84	DAVID	R	JOHNSON		M	DVD_
785	Level 2	00004243	.		08/18/84	DAVID		JOHNSON		M	DVD_
1320	Reference	00001329	1	987315212	08/22/56	JIMMIE	R	GOMEZ		M	DVD_
1320	X-linked	00001713	1	987315212	08/22/56	JAMIE		GOMEZ		M	BOOK
1320	Level 1	00003061	.	987315212	08/22/56	JAMES	R	GOMEZ		M	DVD_

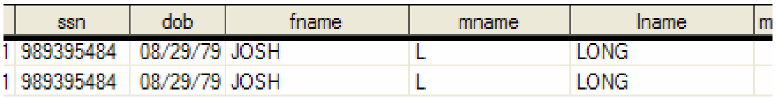
ROYAL ACADEMY: LESSONS LEARNED

As a neophyte to SAS/AF[5] with an overly ambitious vision for this application, I encounter obstacles too numerous to detail in these pages. Two issues and their resolutions, however, are detailed below:

Problem: There is considerable variation in the length of clients' names in large datasets. When displaying a subset of the data in a table viewer, it is an inefficient use of space to have the formatted length of names in the subset default to formatted length of names in the full dataset. Column width may become excessively large. Consequently, the number of variables displayed in the viewer may be reduced, forcing the user to scroll unnecessarily.

Figure 6 illustrates the problem. The length of "Josh" and "Long" is only 4 characters and only a middle initial is present where other records in the dataset had full middle names. There is a lot of wasted space here.

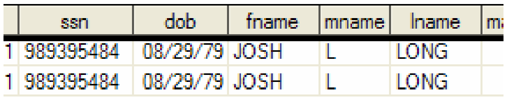
Figure 6



	ssn	dob	fname	mname	lname	m
1	989395484	08/29/79	JOSH	L	LONG	
1	989395484	08/29/79	JOSH	L	LONG	

The data is more efficiently displayed as shown in Figure 7.

Figure 7



	ssn	dob	fname	mname	lname	ma
1	989395484	08/29/79	JOSH	L	LONG	
1	989395484	08/29/79	JOSH	L	LONG	

Solution:

1. Create a macro ("name_format_calc") which:

- determines – within the data subset - the maximum length of a given variable, and
- creates a format for that variable based on the newly determined length :

[5] I embarked on this programming task in April 2004 with no previous experience with SAS/AF or visual basic.

```
%global &format;
%macro name_format_calc (var, format);
  proc sql noprint;
  select max(length(&var))+3 into: form
  from input_dataset;
  %if &form=. %then %let form=1;
  %let form=$%eval(&form.).;
  %let  &&format=&form;
%mend name_format_calc;
```

2. Create a labeled section of SCL ('formats") that:

- Calls the macro "name_format_calc" for each of the string variables being displayed to determine the maximum width of that variable in the subset being displayed in the table viewer. This is done within a submit block.

- Apply the newly identified minimum column width for the subset to the "columns" attribute of the table viewer using the following SCL functions:

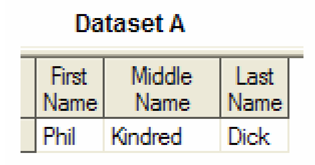
```
objectName._getColumnNumber( columnName, columnNumber )
objectname.columns{columnNumber}.format
```

```
formats:
 sasdataset1.editmode='browse';
 submit continue;
  %name_format_calc(client_first_name, fn_len2);
  %name_format_calc(client_last_name, ln_len2);
 endsubmit;
 sasdataset1._getColumnNumber('client_first_name',columnNumber);
 sasdataset1.columns{columnNumber}.format=symget('fn_len2');
 sasdataset1._getColumnNumber('client_last_name',columnNumber);
 sasdataset1.columns{columnNumber}.format=symget('ln_len2');
 sasdataset1.editmode='tableleveledit';
return;
```

3. "Links" to this labeled section of SCL from the frame's initialization section:

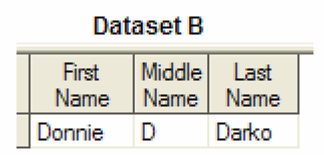
```
init:
  initialization section program code….
  Link formats;
return;
```

Problem: When linking two datasets, part of the process involves concatenation of the two datasets being linked. If the length of string variables (e.g., first name, last name etc.) varies between the two datasets, one cannot simply append one dataset to the other because string fields may be truncated. For example, if dataset A and dataset B are concantenated without reconciling the differing lengths of the name fields, dataset C will result. Notice the Donnie has been truncated to Donn and Darko to Dark.



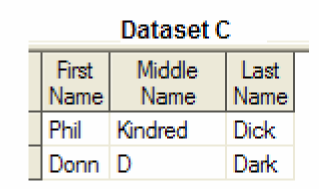
Dataset A

First Name	Middle Name	Last Name
Phil	Kindred	Dick



Dataset B

First Name	Middle Name	Last Name
Donnie	D	Darko



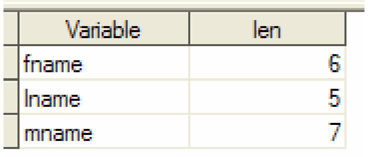
Dataset C

First Name	Middle Name	Last Name
Phil	Kindred	Dick
Donn	D	Dark

A look inside "Presto" the Data Importing and Formatting Wizard's spell book reveals the incantation to vanquish this pesky demon.

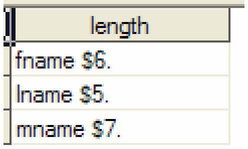
Solution:
Create a macro which - using SAS's ODS functionality – determines where lengths differ for string variables common to both datasets and – using a SAS macro variable – construct an "Length" statement to insure no string variables are truncated during concatenation.

```
%macro concantenate:
  ods listing close; *prevents SAS from sending the subsequent "proc contents" output to the output window";

*creates 2 datasets which contain information about the contents of the datasets being contenated ("matching"
and "sample");
 proc contents data=matching;
 ODS OUTPUT Variablesalpha=m_cont; run;
 proc contents data=sample;
 ODS OUTPUT Variablesalpha=s_cont; run;

*create a dataset which contains – in all instances where the length of the
  2 string variables differs – the name of the variable and the maximum length.
 proc sql;
  create table vars as
  select a.variable,  max(a1.len, ab.len) as len
  from s_cont a, m_cont b where
   a.variable=b.variable and
   a.type='char' and
   a.len ne b.len;
```



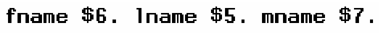
Variable	len
fname	6
lname	5
mname	7

```
*combine the variable name and maximum length to create a  "format style"
  variable;
 data vars;
 set vars;
 keep length;
 length=left(trim(variable))||' $'||(compress(len))||'.';
 proc print;
 run;
```



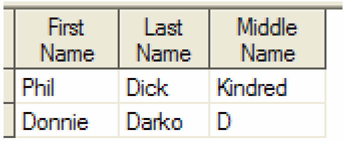
length
fname $6.
lname $5.
mname $7.

```
 *string the format into a single line for use as a length statement;
 proc sql noprint;
 select length into: lens separated by ' ' from vars;
```



```
fname $6. lname $5. mname $7.
```

```
 proc datasets library=work nolist ; delete vars m_cont s_cont; run;
 ods listing;

*use the macro variable with a length statement in a data step;
 data all_ids;
 length &lens;
 set sample  matching;
 run;
%mend concatenate;
```



First Name	Last Name	Middle Name
Phil	Dick	Kindred
Donnie	Darko	D

CONCLUSION

The Link King - powered by SAS/AF - is a state-of-the-art application for record linkage and unduplication. The power of SAS is harnessed to quickly and accurately process extremely large datasets. The Link King has successfully unduplicated 4 million records in less than 6 hours.[6] SAS/AF provides an interface that allows users with minimal SAS programming skills to easily accomplish this inherently complex task.

Rule Your Data !

[6] Using SAS v9.0 with dual 3189 MHz processors.

ACKNOWLEDGMENTS

The Link King gratefully acknowledges:

Dennis Deck (RMC Research, Portland, Oregon), Anton Cox (Oregon Health and Science University, Portland, Oregon), Antoine Krupski and Carole Broderick (Washington State Division of Alcohol and Substance Abuse) for their vision and support.

SAS Technical Support consultants Adam Budlong, Lynn Curley, Ann Rockett, and Kat Sizemore for their patience and expertise.

Gwen Grams and staff at Oregon State Department of Human Services for their invaluable contribution as a test site for The Link King.

Sylvia and Kiki Campbell for their tolerance of my obsession with this project.

Pete Lund (Looking Glass Analytics) for his encouragement in the initial stages of development.

Reid Psaltis for his excellent rendering of The Link King logo.

And all the “early adopters” for their help in identifying bugs.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kevin M Campbell, DrPH
Washington State Division of Alcohol and Substance Abuse
Box 45330
Olympia, WA 98501
Email: thelinkking@yahoo.com
Web: www.the-link-king.com

©