History and Computing 7, 90-103 (1995)

Multi-Source Family Reconstruction

Gerrit Bloothooft

A system for automatic family reconstruction from data from various historical sources is described. A normalized person-oriented data model and normalized data serve as a basis for linkage, while iterative improvement of the linkage structure is made on the basis of constructed reports on individuals.

I Introduction

Record linkage has often been presented in relation to specific projects and is consequently often tailored to the demands of such an application. Although this has led to many insights in the problems of historical record linkage, it hampered comparison of different approaches, and made it hard to develop thoughts on a generalized theory of the linking process. We wanted to study automatic record linkage in the case of information from a great variety of different sources. This may be considered the most difficult, but also the best environment to investigate the process of record linkage in such a way that the resulting approach has a more general validity.

A second challenge is to reduce user-interaction to a minimum. This forces one to think about optimal strategies where user-interaction is concentrated in a few phases in the procedures, while the majority of the linkage processes is realized fully automatic. Such an approach goes against the tendency to develop computer-assisted systems for nominal record linkage[1]. We believe, however, that the boundaries of what can be realized automatically in record linkage have not yet been explored well enough. In general, there will be a relation between the amount of information in sources and the possibilities for automatic reconstructions. If there is limited information, it is unlikely that true links can be made automatically, and consequently, user assistance is needed in an early stage of the process. On the other hand, if there is a web of interrelated information, it may be even possible to consider links under the assumption of serious writing errors. In such a situation the loose description of record linkage, like the investigation whether *John Smith* in one record denotes the same person as *John Smith* in another record, may be widened to the investigation whether *John Smith* in one record shows up as *William Smith* in another record. The latter decision requires very strong evidence, of course; but this evidence is sometimes available, and the required reasoning can be done automatically. An interesting example is given in the appendix. This shows that, in the presence of rich information sources, automatic reasoning can be very powerful and should not be abandoned too quickly for systems that need user interaction. In our view, we may require the computer to generate best guesses of linkage patterns, while the user gets the possibility to disagree with the result and to make improvements manually.

II AN OVERVIEW

The record linkage system described here is called Genesis. Because the system is rather complicated, we first present an overview of its general structure and philosophy. A strong feature is the systems modularity and transparency. This structures the further presentation of details in separate sections. Figure 1 presents an overview of the system.

Of major concern in automatic family reconstruction are the data models and the data processing strategies to be used, which are, of course, interrelated. In a multi-source environment, it is essential to obtain a normalized data model as soon as possible; otherwise,

the linkage procedures will become extremely complicated. But one has to recognize and use the specific characteristic of various sources too. We have solved this by creating a first database that stores information from specific sources in separate tables: the Source Database. This information is transformed by source-specific analysis procedures into a second, independent database with a normalized, relational data model: the Analysis Database. Once the data are brought at this level, all subsequent linking procedures are source-independent. One cannot over-emphasize the importance of the data model of the Analysis Database. This forms the backbone of record linkage procedures which we claim to have a general validity. The Source Database may depend on application, region, period, and so on, but once the data are transformed into the Analysis Database, all special pecularities got rid of before proceeding in the subsequent analyses. Of course, a normalized data model cannot contain all the information that is present in all kinds of sources and this may be a serious limitation. However, as long as this data model can cope with the major information needed for family reconstruction, it may be a very good compromise.

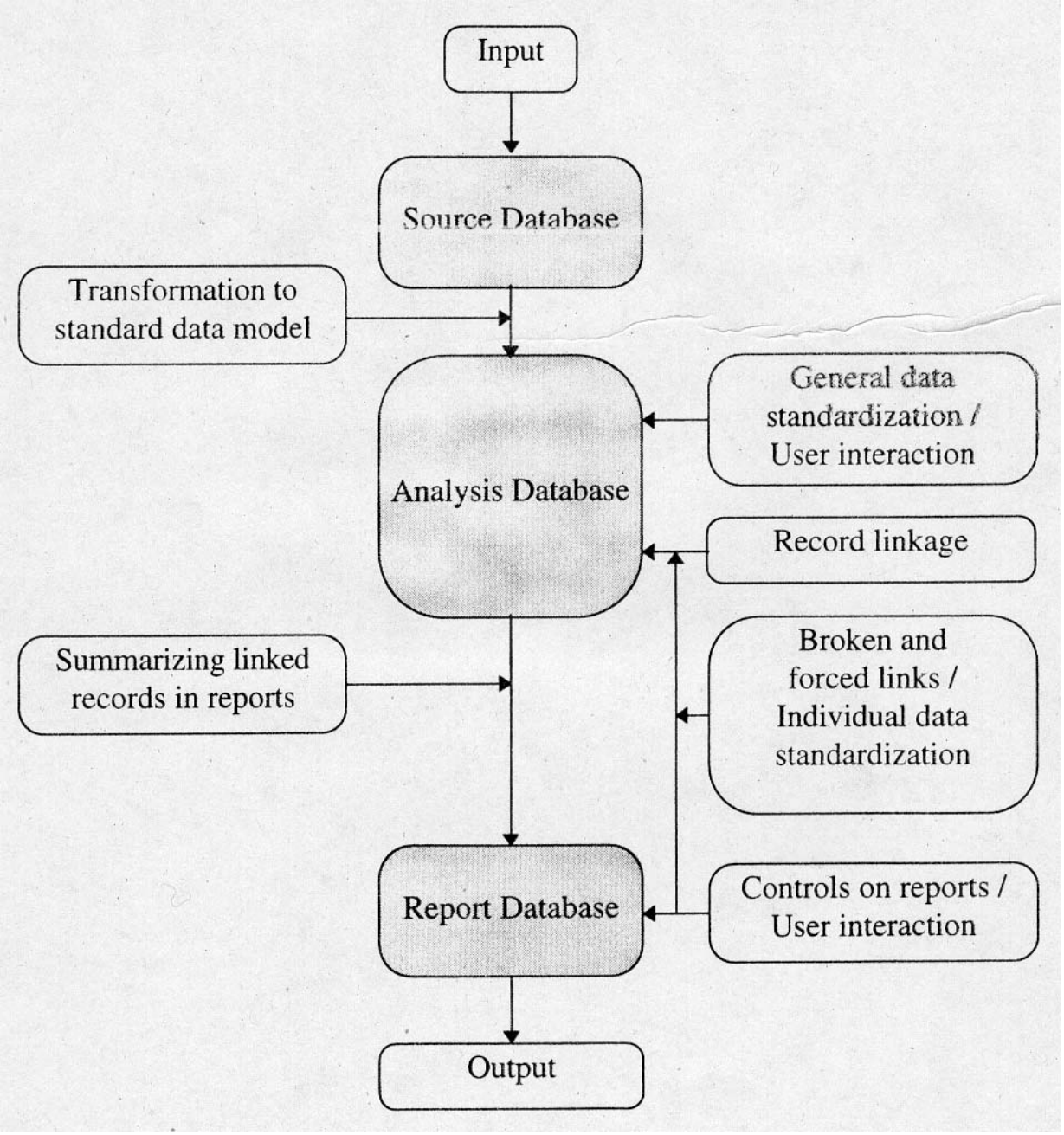
A further preparation for linkage concerns a normalization of the data themselves. Normally, this will be done for nominal data, but in principle this normalization can be applied to all kinds of data types. Another step relates to an estimation of the information stored in each record that can be used for the process of reconstruction. If a record contains little information, it is considered unsuitable for automatic analysis and set aside for user handling. After this preparatory phase, the Analysis Database contains normalized data within a normalized data model.

The linkage process starts with the well-known step of primary links on the basis of matrimonial couples[2]. Next, linkage is performed per data set of equal first name and gender, ordered according to the amount of related information. The procedure starts with the record that has most related information, and therefore probably has the best anchoring in the reconstruction: the target record. Other records in the set are rejected for linking on the basis of inconsistencies with the target record at the nominal level (of the target person, the partners, and the parents) and at the level of ranges for birth, marriage, and decease. The remaining records are subsequently checked for mutual consistency. The linked records are labeled, and the procedure continues with a target record that now contains most related information.

It can be shown that the procedure outlined above does not utilize all information on relations between individuals. The best way to proceed is to summarize the information of the linked records in a report first. For these reports, a third database is needed: the Report Database. Again, it is of great importance to maintain the transparency of the process, and not to merge the report representations with the normalized data representations in the Analysis Database. The Report database uses a relational model comparable to that in the Analysis Database, and summarizes all the information in linked records.

On the basis of reports, several control routines can check the initial reconstruction with respect to relational consistencies between reports and, at the nominal level, with respect to errors or deviations in name standardization or even name changes (the wife that takes the surname of the husband after marriage). This will produce broken links and sometimes forced links at the level of the Analysis Database. This report-based control phase is very powerful in the linkage process.

Figure 1. Schematic overview of the record linkage system: its databases and major procedures.



If the control routines result in changed links, the first linking phase at the level of the Analysis Database is repeated. This relinkage has (and needs to have) the interesting feature of being able to break old links and to create new ones. This feature is also essential for dealing with new data that comes available to the system. One might even argue that the capability of adding new data to a system for record linkage without doing the entire analysis all over again should be a an essential argument in the assessment of such a system.

After relinkage, new reports are made and some old ones are deleted. A new control phase is started, which may result in relinkage, and so on. In practice, a few iterations will lead to a stable result.

Finally, there is the possibility of user interference on the basis of the resulting reports (and the records that have been neglected because they contain too little information). This user interference has the same effect as the control routines and will lead to a (forced) restructuring of the result. Different types of output facilities are present to show the final reports. These may contain all original source information, because links to these data have been carefully preserved during the whole process.

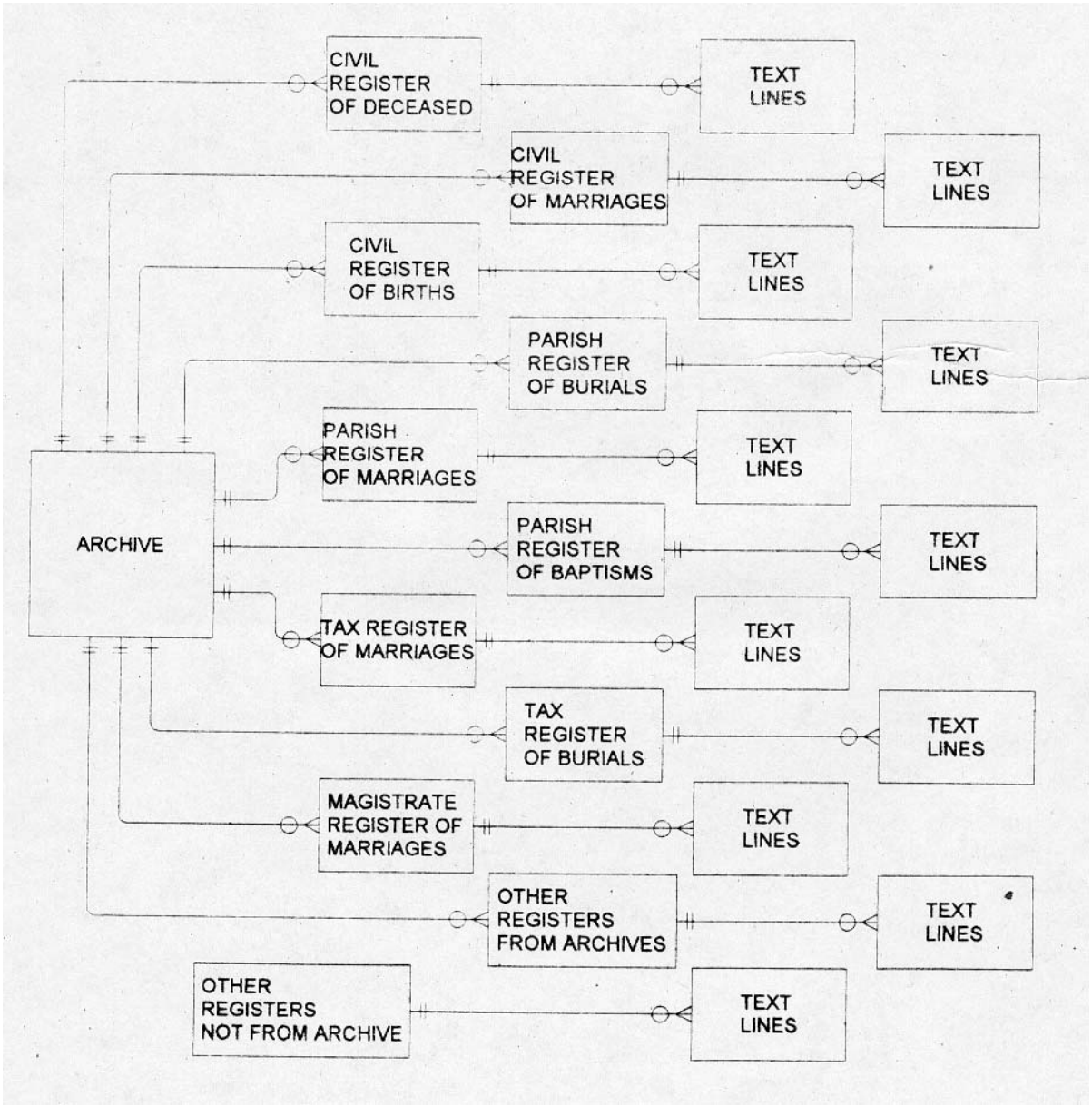
III The Source Database

Standard data models have been developed for the most common genealogical sources in the Netherlands. Sources are the Civil Registration after 1811 (birth, marriage, death), Parish Registrations before 1811 (baptism, marriage, burial), Tax Registrations before 1811 (marriage, burial), and Magistrates' Registrations (marriage) before 1811. Each of these sources are given a separate table (with a flat structure for computational convenience). For other sources of a more variable format, such as notary certificates and relief registrations, one relationally structured table has been developed. There is a separate table 'ARCHIVE' for the description of the specific sources used. Easy input, modification, deletion, and selection options are available. The Source Database is never affected during the linkage procedure. In addition to the input of data in fields according to the data model, there is unlimited space for annotations and original text in separate relational tables. Figure 2 gives the entity-relationship diagram[3] of the Source Database.

Although we used our own specific way of representing source information, there are, in principle, no specific demands with respect to data storage at the source level. In Genesis we used simple dBASE files[4], but more advanced data storage systems with free length fields or even free text storage with tags are conceivable. This does not influence the rationale of the system. Ideally, a user should be able to add new types of sources to the system, or to add new fields. In the present version, Genesis does not have this option, but inclusion of such an option should not cause real problems at the source level. However, complications may arise with the procedures needed to transform these newly added data to the Analysis Database with a normalized data structure.

Names can be imported in Genesis in the original spelling. At this level, the user may have an ambiguity problem with the distinction of the first name(s), the patronymic form (genitive) and the surname[5]. Later on in the linkage process, checks are carried out to determine whether other interpretations of parts of a name lead to better results. The given problem of interpretation of information is a general one, however, and will be encountered in any type of data storage.

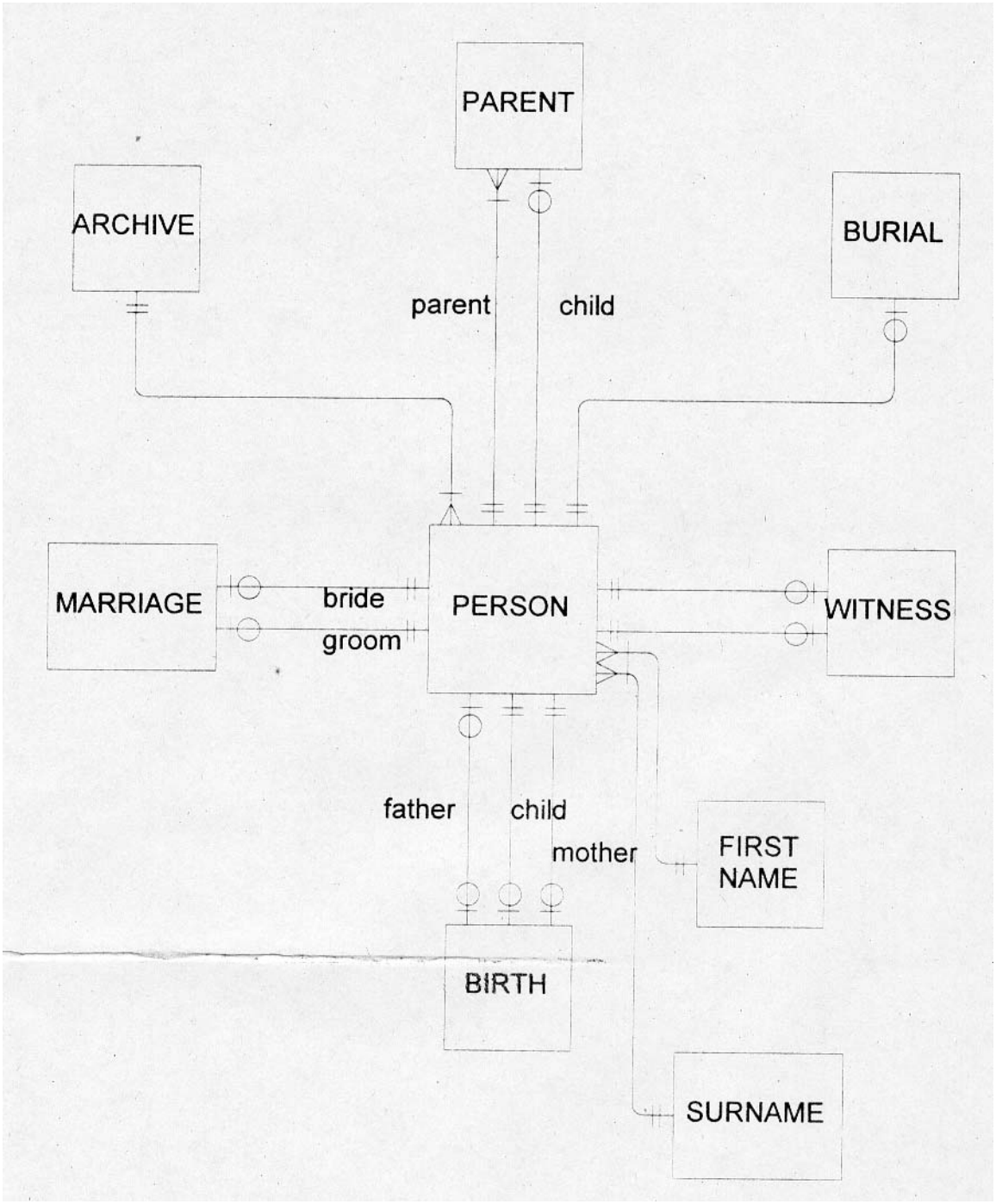
Figure 2. Entity-Relationship diagram of the Source Database of Genesis. All tables have a fixed, flat structure, with the exception of the OTHER REGISTERS FROM ARCHIVES table, which is relational and lists persons and their family relations mentioned in any certificate.



IV The Analysis Database

The normalized data model of the Analysis Database is of vital importance for the linkage procedure. We have opted for the relational model with tables named PERSON (for nominal and other general information), BIRTH (also including baptism), MARRIAGE, BURIAL (also including death), PARENT, and WITNESS (for other persons that have been present or mentioned during registration of birth, marriage, death, or any other event mentioned in the sources). Figure 3 gives the entity-relationship diagram of the Analysis Database.

Figure 3. Entity-Relationship diagram of the Analysis Database of Genesis. The ARCHIVE table is the same is in figure 1 and present the link with the original source data. FIRSTNAME and SURNAME tables are used for the purpose of name standardization.



The chosen data model allows to accommodate most of the genealogical information present in the various sources. The transformation of the original data into this model involves a reasoning on what can be deducted from the original data with respect to birth, marriage, and decease. The major feature here is the introduction of ranges in time during which some event should have happened. From the registration of the birth of a child, for instance, it follows that the mother is between a minimum and maximum age (or age is given), while she will die between the child's birth and before she reaches some maximum age. For the father some (more loose) ranges for birth and death can be estimated also. This type of ranges is applied for birth, decease and marriage (ranges for marriage date and for end-of-marriage date). The

transformation is governed by utilization of implicit knowledge on relations between persons mentioned in some source, and eight parameter values that can be set by the user, presented in Table 1.

Table 1. Parameters for the transformation of source data into the normalized structure of the Analysis database. Default values are given, but can be changed.

Minimum age for occupation	6 years
Minimum age at child birth	17 years
Minimum age for a witness	20 years
Maximum age at first marriage	60 years
Maximum age	100 years
Maximum age difference between partners	30 years
Maximum age at child birth for mother	50 years
Maximum age at child birth for father	70 years

These parameter settings are closely related to nine rules given by Wrighley and Schofield[6]. It is important to note that at this stage we do not create new knowledge, but that we try to draw conclusions from the information given in a source. This is essentially different from the Wrighly and Schofield approach.

As an example, we give an analysis of the following simplified birth registration:

William, son of John and Mary, was born in 1768

The system will make the following derivations and will put these in the tables of the Analysis Database:

William was born in 1768
William died between 1768 and 1868
William's parents were John and Mary

John was born between 1698 and 1751
John was married to Mary before 1767
John was married to Mary until between 1767 and 1851
John got a son William in 1768
John died between 1767 and 1851

Mary was born between 1718 and 1751
Mary was married to John before 1767
Mary was married to John until between 1767 and 1851
Mary got a son William in 1768
Mary died between 1768 and 1851

The derivations result in an explicit description of what is known from the viewpoint of each mentioned person. In other words, the description is fully person-oriented. Such a strategy results in redundant data storage but this proves to be extremely useful in the further process of linking.

In the present version of our system, the transformation of data (with the exception of the parameter setting) is made in such a way that it cannot be influenced by the user. This may be specially problematic if one wishes to add new source tables or new fields to the Source Database. Ideally, it should be possible that the user can describe the source data and their relations in such a way that the necessary transformation to the normalized tables can be made automatically.

Another limitation relates to the absolute values of the parameters. If one wishes to be sure never to make an error, the settings should be very tolerant, which can give (too) much freedom during the linkage phase. As an alternative, probability distributions could be used, which would only slightly complicate the further linking process.

IV.1 Name standardization

We have to cope with the problem of large variations in the spelling of names (first names, patronymics, surnames and toponyms), related to pronunciation differences, spelling inconsistencies, the appearance of diminutive and latinized forms of names, but also to plain writing, reading, and typing errors. We have developed an algorithm (partly rule-based, partly on a probabilistic basis) that automatically tries to find the best standard for each name[7]. Nonetheless, user control and interaction is absolutely necessary at this level. The names in the PERSON table are transformed into new fields in a standardized form. It should be realized that a single standard for each name can not cope with all spelling variation. A name can have more than one acceptable standard, whereas the choice of standards is often disputable and dependent on historical period and region. Other problems like name changes (wife takes the surname of the husband), confused names, or completely erroneous names can never be solved by spelling standardization. In our approach, we make the best guess for a standard and use this standard in the first linkage round. Later, we reconsider the resulting linkage structure on the possible presence of serious misspellings or name errors. We then introduce new name standards whenever enough evidence is available in individual cases. In our view, this is more efficient than a system in which the name standardization is an intrinsic part of the linkage process itself, where all kinds of alternatives are kept open and considered in comparisons.

IV.2 Information content of a record

As a second, preparatory step we assign a number to each record that exemplifies the expected importance for the reconstruction. The rationale of this is that we prefer, for instance, to start linkage with information that comes from a marriage in a civil register, because a civil register is likely to have much information on ages, names of parents and so on. To realize this, values quantifying the richness of information related to each individual person mentioned are computed (information content). This is done on the basis of more or less arbitrary (implicit) rules[8]. We adopt a minimum level of information content that should be present for a name for further analysis[9]. If the information content does not reach a threshold, it is up to the user whether to use the data for linkage. The definition and application of information content relates to notions such as hierarchical and preferential scoring[10].

IV.3 Strong links

We start to make the strong links. These are the links that can be made on the basis of matrimonial couples. There are some rules (parameters can be set by the user) that specify the necessity of having knowledge on parts of the proper names of the husband and wife. Only under some conditions (with regard to place and time) are proper names that consist of a first name plus patronymic only considered acceptable for a strong link. Experience with various Dutch 18th-19th century material shows that roughly one-third of all links are of this strong type.

IV.4 The general linkage procedure

To arrive at a convergent linkage procedure we use three basic premises:

- the reconstruction should use all available data
- the reconstruction should not be contradictory

- the reconstruction should arrive at a minimum number of persons

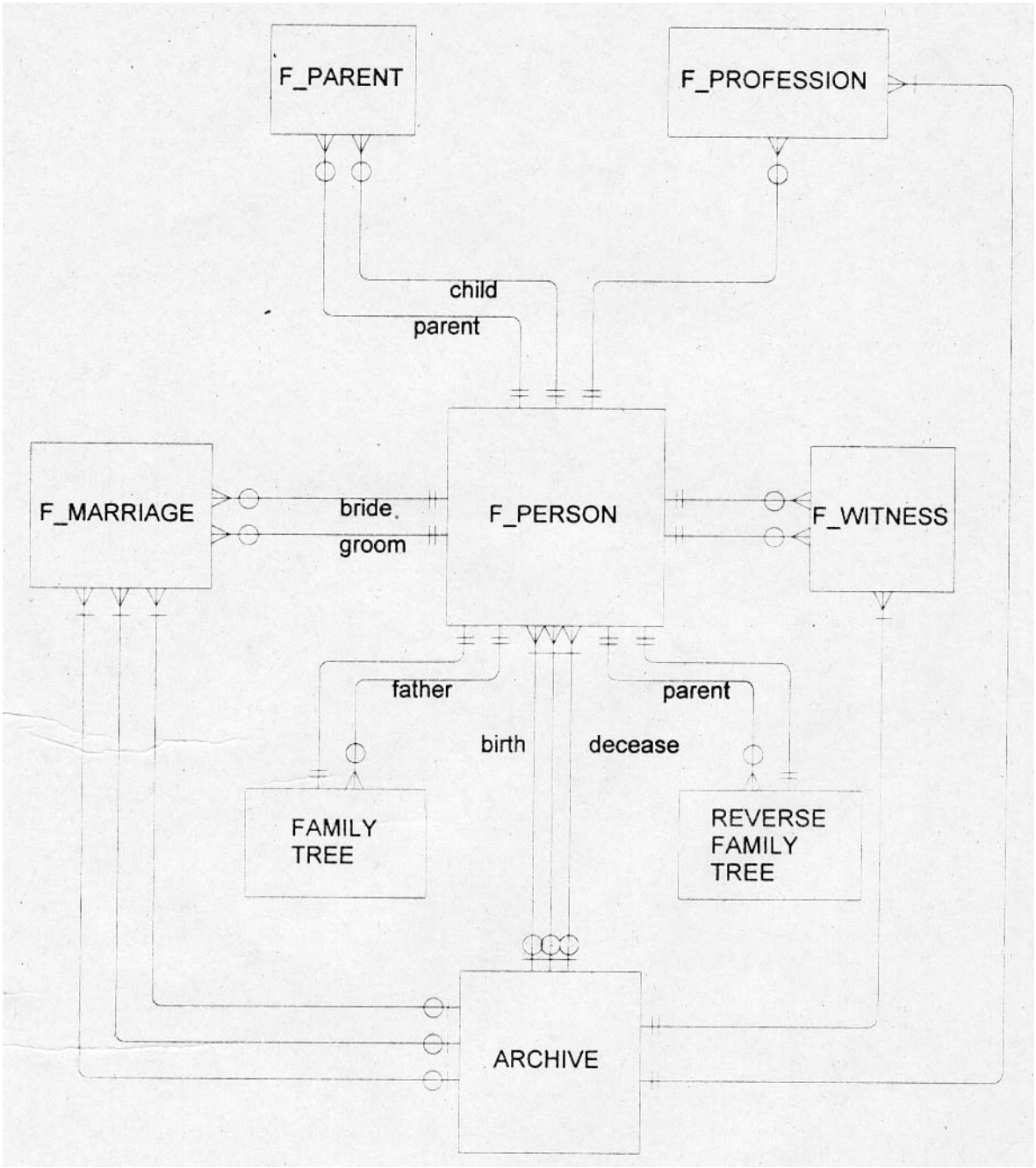
The last premise can be disputed. However, we think this is a good starting point for a convergent solution. In any case, if we allow user interaction to adapt the final solution it certainly is an acceptable premise. Its implication is that we always link records as long as no contradictions arise. Our strategy to choose target records in the order of the amount of related information probably helps to fulfil the premise.

The record linkage is performed on data that are 'pocketed' by (standardized) *first nam*e and gender. The argument to group on first name is that surnames are frequently missing in Dutch material from before 1811. There is no principle deviation from the following procedure, however, if the data are grouped by surname. Within a set of equal first name and gender, we start the linkage with the record that has the highest information content and is most critical to match. This target record is compared to all other records in the set and these are rejected for linking if there are inconsistencies with the target record. These inconsistencies may concern the name of the target person (patronymic and surname), and the full names of present and previous partners, and parents. Furthermore all intervals of birth, marriage and decease are compared and these should overlap the intervals of the target record for acceptance. All records that survive these comparisons are considered to be in agreement with the target record. Still, they may be mutually inconsistent. At this point we may be forced to make a few decisions that are disputable, but which can be improved at a later stage. In the case of interval contradictions between records, we choose for the record with the latest birth, the earliest marriage or the earliest decease. For example, if we know that the target person is born before 1760 and we have two parish baptism registrations, under the same name, in 1755 and 1757 that are both consistent with the target record but that are obviously mutually inconsistent, we opt for the latter. In the end, all records that are consistent with the target record are labeled, and the procedure continues with a target record that now contains most related information.

In general, records with little information are consistent with many other records, but with a high probability of an erroneous link. These records constitute the major problem in linking. Records with a lot of internal information are more critical in comparisons with other information-rich records. The procedure we have chosen results in a situation where the record with the most information has a preferred position to attract records with little information. This is arbitrary and apt to errors, but any other solution will have the same drawback. Our approach has the advantage that it is well-defined and that it brings order in otherwise chaotic attempts to make links. Also for the relinkage process following the control phase, the chosen order limits the complexity of the process.

Another limitation in the present version of Genesis is that we do not use place of residence and occupation because of the great variation found in this type of information. We feel that such uncertain information cannot be decisive in the forced judgement whether records are consistent or not. It can be argued, however, that in situations where only nominal data, occupation and place of residence are known, this information cannot be neglected and may play a vital role in the linking process. Also in the case of decisions like 'latest birth, earliest marriage and earliest decease', one may weigh place of residence and occupation too. Nevertheless, the best stage to consider this less reliable information is the control phase, in which the first overview of linkage patterns is already available and this 'weak' information can be interpreted within the whole frame work.

Figure 4. Entity-Relationship diagram of the Report Database of Genesis. The ARCHIVE table again gives the link to the original source data. Furthermore, the F_PERSON table is linked to the PERSON table from the Analysis Database (Figure 3). The FAMILY TREE and REVERSE FAMILY TREE tables are not used in the linkage process but facilitate the building of family reconstructions afterwards.



V The Report Database

The initially linked records do not comprise the final linkage result. The main point is that we did not use all knowledge on relations between persons, because the linking procedure handles persons separately. But before we can use knowledge on relations to our advantage,

we have to make reports that summarize all the linked information in records per person. For instance, we may know that two persons have the same parents. If they have the same first name we may assume the older to be deceased before the birth of the younger. For the system, this is new information that is only available if we have a report on each of the parents in which their children are mentioned.

In the Report Database every person is mentioned only once. The relational tables have a structure that compares the analysis tables and are named F_PERSON, F_MARRIAGE, F_WITNESS, F_OCCUPATION, and F_PARENT, see Fig. 4. An important difference with the analysis tables is that F_PERSON includes data on birth and decease, since these events can only happen once. F_PARENT is redundant to the birth information in F_PERSON and is created for computational reasons only. Another important difference with the Analysis Database is that in each of the report tables there is a (sometimes multiple) reference to the Archive table for each record, whereas in the Analysis Database only the PERSON table has a reference to ARCHIVE. This is because a record in the PERSON table refers to a single source event, while the report tables summarize information from various events in different sources.

V.1 Control routines

Once the reports are available, a number of control routines come into action. First, there is a check on reports that mention more than one marriage for the person involved. This may be reality, but it may also be the result of incorrect name standardization or name errors in the original registers (we give an example in the appendix). Therefore, we make name comparisons with looser boundary conditions for both first names and surnames of partners, or even bring highly different names under the same standard if there is enough evidence to do so. A very difficult problem arises when a wife adopts the surname of her husband. At the report level we may look for this possibility and adapt the standardized name, for this special occasion only. The same holds for cases where no first name have been mentioned but only the status 'widow'. Another interesting control possibility is that a previous partner should have been deceased before the next marriage of a partner (Genesis does not recognize a divorce yet!). If it is explicitly known that the earlier wife still lived after the next marriage, we have made an erroneous link. Another control is on birth and decease of children of the same name. In all cases relinkage is necessary.

We give two examples. In the first example we have two sons of the same name.

report on William1:
William, son of John and Mary, was born in 1768
William died in 1780

report on William2:
William, son of John and Mary, was born in 1773

report on John:
John was married to Mary before 1768 until after 1772
a son William was born in 1768
a son William was born in 1773

report on Mary:
Mary was married to John before 1768 until after 1772
a son William was born in 1768
a son William was born in 1773

The reports of the parents learn that William1, born in 1768, logically should have died before 1773, and not in 1780. This new evidence is added to the data and after relinkage we will see the following reports on both Williams:

report on William1:
William, son of John and Mary, was born in 1768
William died before 1773

report on William2:
William, son of John and Mary, was born in 1773
William died in 1780

The second example may arise when at the death of a wife the husband has not been mentioned in the register. Without this knowledge, a link implying a second marriage is not impossible from the viewpoint of the man.

report on John:
John and Mary were married in 1755
John and Elizabeth were married in 1770

report on Mary:
Mary and John were married in 1755
Mary died in 1780

The comparison of the reports of John and Mary learns that (a) either John did not marry a second wife Elizabeth, or (b) Mary who married John did not die in 1780, but before 1770. Without any other evidence, the (arbitrary) preference will be given to the first possibility, resulting in a broken link with John who married Elizabeth in 1770.

These are very simplified examples, and records with so little information normally will be discarded for linkage. However, they nicely illustrate some features of the linking process we adopt and the iterating approach to converge to a consistent solution. It would be very difficult to realize the same solution in one run only (if not impossible, as for the example given in the appendix).

After these controls an update of missing surnames follows (both from father to children as from children to father). This is a somewhat risky procedure in case of erroneous links. Finally, reports are submitted to an optimization of date ranges. Because of unpredictable effects of relinkage, the entire control procedure is run a few times to arrive at convergence.

V.2 Relinkage

As has been said earlier, the power of relinkage (or the handling of new data) is an important feature of an efficient record linkage system. Suppose we have new information, then there are three options:

- the additional information relates to a new person and does not influence the existing reconstruction
- the additional information adds to an existing report and is not in conflict with the existing reconstruction
- the additional information results in a new interpretation of the existing reconstruction and breaks old links (and existing reports)

The first two cases do not present any problem, but the third one may lead to an avalanche of changes in links that should be treated carefully. If broken reports and links occur during relinkage, *all* records of the broken reports are again considered to be new records that have to be subjected to relinkage. All broken reports are marked for deletion and new reports will be made on the basis of the new links. The advantage of this approach is that changes in the family reconstructions are made locally. We feel that this compares to human approaches of the problem.

Control routines and the iterating process of controls and relinkage are of eminent importance to arrive at optimal reconstructions. Definition of the control routines and the setting of their parameters is yet beyond the influence of the user. This is an aspect that needs some careful attention in the future. It has been very rewarding, however, to see that many difficult problems in linkage could be resolved in a general way with the aid of control routines due to the transparent structure of Genesis.

VI Output and interactive updates

Genesis provides the opportunity to export reports to screen, printer or file. The reports may be presented in a condensed form, but may also include the original source information the report is based on. Reports may be combined automatically to yield family tree structures. This provides the opportunity to check the results of Genesis easily.

As has been said, some information is not processed if the information content is too low. In such a case user interaction comes into play again. If for these types of records only one report candidate is available, the user may choose to make an automatic link. If more than one report is in agreement with the low-information record, the user can make a decision on the basis of the reports.

In the same way as with low-information records, user interaction is presently under development to adapt the reconstruction derived by Genesis in case of erroneous results, by creating fixed links.

VII Discussion

We do not claim by any means that we have presented the basis of the ultimate family reconstruction system. Nevertheless, we have shown that the structure of three layers of databases, combined with a normalization of both data model and the data itself provide an excellent basis for a transparent system. We have mentioned various improvements and directions of further development that should have our attention in the near future. We find it very encouraging that the present system already yields good results with difficult data[11]. Of course, human judgement will always be necessary for making a final reconstruction. We can make systems according to laws of logic and probability that may be so powerful that even partly erroneous data can sometimes be interpreted very reasonably, but computers can never be taught to model all the surprising and unpredictable reflections of real life.

Notes

[1] See the articles 'Matchmaker, Matchmaker, Make Me a Match' by J. Atack, F. Bateman, and M. Eschelbach Gregson, *Historical Methods* 25 (1992), pp. 53-65, and 'Computer-Assisted Record Linkage Using a Relational Database System' by.J.E. Vetter, J.R. Gonzalez, and M.P. Gutman, *History and Computing* **4** (1992), pp. 34-51.

[2] G. Bouchard and C. Pouyez, 'Name Variations and Computerized Record Linkage', *Historical Methods*, 13 (1980), pp. 119-125.

[3] A description of the technique of making entity-relationship diagrams can be found in J. Martin, *Recommended diagramming standards for analyst and programmers: A basis for automation* (Prentice Hall, 1987). The author is indebted to Toine Schijvenaars for making the three diagrams presented in this paper.

[4] Genesis has been written in Quicksilver and all data files have a dBASE format. Every phase in the reconstruction has its own, compiled programme, while the calls to these programmes are made from a menu shell. Genesis runs on a AT 486 DX II (66 MHz). For relatively large applications (20.000 records in PERSON), total analysis duration was about 24 hours. The

total computation time is about linear with the number of records in PERSON, which amounts to a raw average of about 4 seconds per name.

[5] For example, in the name *Jan Dirck Janse*, *Jan* but also *Jan Dirck* may be interpreted as first name(s), *Dirck* but also *Janse* may be a patronymic form, while *Janse* can be the surname too.

[6] E.A. Wrighley and R.S. Schofield, 'Nominal record linkage by computer and the logic of family reconstruction', in: E.A. Wrighley, (Ed), *Identifying People in the Past* (London, 1973), pp. 64-101.

[7] Our procedures for name standardization have been described in G. Bloothooft, 'Corpus-based name standardization', *History and Computing*, 6 (1995), pp. 153-167.

[8] We assign points to information present in a record that relates to the individual concerned: first name of person, partner, parent (one point); patronymic of person (if father is not mentioned), partner, parent (three points); surname of person, partner, mother (five points); any date interval of birth and death less than two years (five points), between two and five years (three points), between five and ten years (two points), more than ten years (one point); date of marriage (three points).

[9] The minimum amount of points needed is usually three, but this normally leads to too much links. A threshold between five and ten is safer, but excludes more records from the linkage process, depending on the application.

[10] E.A. Wrighley and R.S. Schofield, 'Nominal record linkage'; G. Bouchard, 'Current Issues and New Prospects for Computerized Record Linkage in the Province of Québec', *Historical Methods* 25 (1992), pp. 67-73.

[11] We have tried Genesis on various 18th and 19th century datasets. In the case of family archives, with pre-selected data that are strongly interrelated but with frequent missing surnames, almost perfect reconstructions could be realized. In another project the study of poor families in a rural village (1770-1810) was undertaken on the basis of electronic versions of full parish registers, relief registration, and tax registers. Surnames were always available (but see the appendix!). Main problems here were the relative short period, 40 years, that in general did not cover a lifetime, (resulting in overlinkage) and serious errors in dates of the poor relief register (resulting in missed links). Some overlinking could have been avoided if capital tax and place of residence were used in the analyses. Still, the reports Genesis produced enormously facilitated the reconstruction of the family histories of the poor.

Appendix

Example of the effect of control routines on reports

The example below nicely shows the effect of control routines on names and patronymics on the basis of an initial report of a man, *Gerardus Thomas Verschuren*, that contained an initial hypothesis of four marriages. The surnames *Verschuuren*, *Verschuren*, *Verschuure*, and *Verscheuren* already got the same standard in the name standardization, as did *Swenkel*, *Swinckels*, and *Swinkels*. We only reproduce the part of the report on the marriages and do not give further details on children. The first and second marriages are known from the Magistrate marriage registration, the third and fourth originate from the baptism registration of children. Note that the ranges of duration of each marriage are not yet tuned, but computed on the basis of the original data. Adaptation of these ranges is part of the control phase. Original name spellings are given in italics.

Report after the first linking phase:

Gerardus Thomas Verschuren

Married 1. on 07-05-1780 in Aarle-Rixtel, living in Aarle-Rixtel, until between 1780 and 1784 with Maria **Cornelis Verstappen** (4673) from Aarle-Rixtel
<*gerit thomas verschuuren and maria cornelis verstappen*, marriage for magistrate 1780>
<*gerardus verschuren and maria verstappen*, baptism child 1781>
<*gerit verschuure, widower of maria cornelis verstappen, and adriana goord swinkels*, marriage for magistrate 1784>

Married 2. as a widower on 23-05-1784 in Aarle-Rixtel, living in Aarle-Rixtel, until between 1784 and 1863 with **Adriana Goord Swinkels** (4727) from Aarle-Rixtel.
<*gerit verschuure, widower of maria cornelis verstappen, and adriana goord swinkels*, marriage for magistrate 1784>

Married 3. between 1744 and 1785 until between 1797 and 1863 with **Arnolda Godefridus Swinckels** (2492)
<*gerardus verschuren and arnolda swenkels*, baptism child 1785>
<*gerardus verscheuren and arnolda swinckels*, baptism child 1790>
<*gerardus verscheuren and arnolda swinckels*, baptism child 1792>
<*gerardus verschuren and arnolda godefridus swinckels*, baptism child 1796>
<*gerardus verschuren and arnolda swinkels*, baptism child 1797>

Married 4. between 1744 and 1787 until between 1787 and 1863 with **Godefrida Swenkels** (4121)
<*gerardus thomas verschuren and godefrida swenkels*, baptism child 1787>

Automatic control routines consider these data and adapt name standards for individual cases. After that follows relinkage, and the report given below. Original spellings of names are only given for the first marriage, because for that marriage extra evidence has been found. Again no details on children are presented.

Gerardus Thomas Verschuren

Married 1. on 07-05-1780 in Aarle-Rixtel, living in Aarle-Rixtel, until 1783 with **Maria Cornelis Verstappen** (6612), from Aarle-Rixtel
<*godefridus verschuren and maria verstappen*, marriage parish 1780>
<*gerit thomas verschuuren and maria cornelis verstappen*, marriage for magistrate 1780>
<*gerardus verschuren and maria verstappen*, baptism child 1781>
<*gerrit thomas verschuuren and maria corn. jan verstappen*, burial parish 1783>
<*gerit verschuure, wednr van maria cornelis verstappen and adriana goord swinkels*, marriage for magistrate 1784>

Married 2. as a widower on 23-05-1784 in Aarle-Rixtel, living in Aarle-Rixtel, until between 1797 and 1863 with **Arnolda Godefridus Swenkels** (4727), from Aarle-Rixtel

For the first marriage there is extra evidence from the parish marriage registration because the system concluded that the first names *Godefridus* and *Gerardus* (with variants *Gerit, Gerrit*), that are etymologically entirely different, imply the same man. The death of the wife

could also be included, because the system corrected the erroneous interpretation of the patronymic in the burial register: the patronymic is *Corn. Jan* (= *Cornelis Jan*) and not *Jan* (with first names *Maria Cornelia*). This incorrect user interpretation of the abbreviated name blocked the correct linkage in the first phase.

There are three different first names for the second wife: *Adriana*, *Arnolda*, and *Godefrida*. The patronymics *Goord* and *Godefridus* have the same internal standard and are equivalent. In such a case, there is enough evidence for Genesis to combine the three first names and to standardize these to the most frequent name: *Arnolda*.

The author

Gerrit Bloothooft is a staff member of the Department of Computer & Humanities of Utrecht University, The Netherlands. He took his Masters in Technical Physics, his PhD on 'Spectrum and Timbre of the Singing Voice', and is presently responsible for the curriculum specialization in Speech Technology. He transferred his knowledge of automatic speech recognition techniques to the field of name standardization and historical record linkage.

Not in the paper:

Abstract
The complex problem of automatic family reconstruction on the basis of multiple historical sources can only be solved within the framework of a transparent and modular system. The source information (gathered from multiple sources in various tables in the Source Database) should be transformed and normalized in such a way that a source-independent representation is realized (in a second, independent Analysis Database). It will be argued that this source-independent representation provides an excellent basis for defining common grounds between record linkage applications that differ widely in historical period and region. Records are pocketed on the basis of the first name. The subsequent linking process (link-by-link) is governed by the information content of a person in a record and uses first name, patronymic and surname, names of parents and partners, date intervals for birth, marriage, and death, and residential information in some cases. After the linkage stage, a report summarizes all available information per individual, for which a third, independent Report Database is available. On the basis of these reports, automatic control routines improve the linkage result iteratively. User intervention is allowed at the stage of name standardization and in a final stage, checking person reports and initiating relinkage. The system has the powerful property to allow for local updates of the links (with respect to new data and during relinkage).