Proposte e ricerche. Rivista di storia economica e sociale /
An Italian Journal of Social and Economic History,
anno XLVIII, n. 94 (2025), pp. 151-176, © eum 2025
ISSN 0392-1794 / ISBN 979-12-5704-046-8
DOI 10.48219/PR_0392179494_009

Carlo Anselmi*

Sage: un software per la ricostruzione famigliare

Abstract. L'articolo descrive il software Sage, sviluppato dall'autore, che è utilizzato per la ricostruzione della rete famigliare della popolazione della comunità di Marciana nell'isola d'Elba in un arco temporale di oltre tre secoli, a partire dai registri parrocchiali. Esso è costituito da un *costruttore* che genera un database sostanzialmente compatibile con lo standard Gedcom; un programma *unificatore*, che identifica le possibili occorrenze della stessa persona in diverse citazioni documentali, con un criterio di tipo probabilistico basato su punteggio, e un programma di verifica finale che controlla la presenza di anomalie nel database. L'approccio è definito di tipo *iterativo*, in quanto si alternano cicli di elaborazione autonoma e successivi interventi manuali per la revisione dei risultati e la correzione di eventuali errori. L'autore si propone di rendere disponibile il software per la comunità dei ricercatori.

Parole chiave. *Family linkage*, analisi nominativa, demografia storica.

Sage: A Software for Family Linkage

Abstract. This paper describes the Sage software, developed by the author, which is used to reconstruct the family network of the population of the Marciana community in the island of Elba over a period of more than three centuries, starting from parish records. It consists of a *builder* program that generates a database substantially compatible with the Gedcom standard; a *unifier* program that identifies possible occurrences of the same person in different documentary citations with a probabilistic score-based criterion, and a final verification program that checks for anomalies in the database. The approach is defined as *iterative*, as it alternates cycles of autonomous processing and subsequent manual interventions to review the results and correct any errors. The autor aims to make the software available to the researchers' community.

Keywords. Family Linkage, Nominative Analysis, Historical Demography.

* Corresponding author: Carlo Anselmi (independent scholar), e-mail: carlo.anselmi2105@gmail.com.

1. *Introduzione*. Da tempo le tecniche di ricostruzione famigliare, o *family linkage*, hanno guadagnato un crescente interesse nella comunità degli studiosi di demografia storica, a causa della messe di informazioni che è possibile ricavare dallo studio delle complesse reti di parentela che si ottengono con tali metodologie. Tuttavia, tali risultati vengono raggiunti solo a prezzo di un lavoro enorme. Ciò appare scontato se si considera che L. Henry e M. Fleury, che negli anni Cinquanta del secolo scorso furono gli iniziatori di questa metodologia, lavoravano manualmente. Ma anche oggi, periodo nel quale gli storici hanno a disposizione strumenti informatici molto sofisticati, ricostruire le reti famigliari plurisecolari di un'intera comunità rimane un compito formidabile e in qualche misura elusivo[1].

Negli anni sono stati sviluppati molti programmi che sarebbe impossibile anche solo citare in questa sede. Di fatto essi corrispondono a differenti approcci, talvolta antitetici, con cui i diversi gruppi di ricerca affrontano il problema. L'aspetto chiave che li differenzia è il modo in cui viene risolto il problema dell'unificazione, ossia la decisione di considerare come appartenenti a una stessa persona due riferimenti presenti in documenti diversi.

A un estremo dello spettro possiamo collocare il metodo interattivo, ossia quello in cui l'unificazione viene decisa caso per caso personalmente dallo storico, sulla base degli elementi che gli vengono proposti da un opportuno software[2]. Questo software ha il compito sia di interfacciarsi con il sottostante database, sia quello di curare tutti i dettagli pratici del processo di unificazione. All'estremo opposto troviamo i metodi automatici che, in base a un insieme di regole stabilite a priori dallo storico, decidono autonomamente se procedere o no all'unificazione. Il punto di forza del primo metodo è indubbiamente la qualità dei risultati ottenuti, perché lo storico può sfruttare tutta la sua esperienza e il suo intuito per decidere; d'altro canto, rispetto al secondo metodo, il tempo richiesto è enormemente superiore[3]. Il metodo automatico in generale produce risultati affetti da un maggior numero di errori sia di tipo 1 (mancata unificazione di due registrazioni riferibili alla stessa persona) che di tipo 2 (indebita unificazione di due persone che sono diverse). Tuttavia, secondo alcuni autori la qualità dei risultati ottenuti dipende molto più da quella dei dati di origine, in termini di completezza e precisione, che dalla metodologia utilizzata.

[1] Si veda per esempio M. Gasperoni, *Ricostruire e analizzare un'intera popolazione - Prospettive metodologiche, euristiche e uso del computer*, in «Popolazione e storia», 2010, 1, pp. 73-83.

[2] Si veda per esempio E. Fure, *Interactive Record Linkage. The Cumulative Construction of Life Courses*, in «Demographic Research», n. 3 (2000), art. 11.

[3] C'è da attendersi che, con l'ingresso dei sistemi AI e *deep learning* anche in questo specifico campo di ricerca, molte cose cambieranno e assisteremo in qualche modo a una chiusura del cerchio...

Nella comunità dei ricercatori di demografia storica la discussione su quale sia la strategia migliore per l'unificazione è molto accesa e non mancano le affermazioni perentorie, come quella di R. Schofield che è favorevole a un approccio totalmente automatico: «se il giudizio di uno storico ha una qualsiasi pretesa di rispettabilità intellettuale, i principi sui quali è basata devono essere codificabili in una forma algoritmica e quindi devono essere eseguibili da un computer senza ulteriore intervento umano»[4]. Come vedremo nel seguito, l'approccio da me utilizzato è in qualche modo intermedio e lo scopo di questo lavoro è quello di descrivere sommariamente come è stato applicato allo studio della popolazione di Marciana.

2. *Il software Sage*. Sage è stato sviluppato nel corso di molti anni ed ha avuto una costante evoluzione per adattarsi alle esigenze della ricerca. Attualmente esso è costituito da una serie di moduli che gestiscono le varie fasi dell'elaborazione; nel seguito vedremo in particolare il modulo costruttore e il modulo unificatore.

Qualunque programma informatico richiede che i dati su cui deve lavorare siano predisposti in una forma opportuna. Nel nostro caso ciò avviene in due fasi successive: dapprima i documenti digitalizzati vengono decifrati manualmente e le informazioni rilevanti trasferite in un archivio strutturato, cioè un foglio elettronico tipo Excel. Successivamente un modulo costruttore elabora le registrazioni presenti nel foglio elettronico e, a partire da esse, crea il database finale che è costituito da un file testuale in formato Gedcom[5].

Occorre a questo punto una riflessione su cosa effettivamente troveremo nel database dopo queste prime fasi del lavoro; questo ci porterà a capire cosa si intende per "unificazione" e perché essa ha un'importanza fondamentale nella ricostruzione familiare. Prendiamo in considerazione un caso tipico, cioè la registrazione di un matrimonio; le informazioni che troveremo solitamente sono le seguenti[6]: data, nome sposo, padre sposo, cognome sposo,

[4] R. Schofield, *Automatic Family Reconstitution. The Cambridge Experience*, in «Historical Methods. A Journal of Quantitative and Interdisciplinary History», 25, 1992, 2, p. 75.

[5] Gedcom è l'acronimo di *Genealogical data communication* e rappresenta lo standard internazionale di fatto per la conservazione e trasmissione dei dati genealogici e fu inizialmente sviluppato dalla Chiesa di cristo degli ultimi giorni nel 1984. L'aspetto essenziale di questo formalismo è che esso è in grado di rappresentare sia le informazioni di carattere "orizzontale", quali nome, cognome, data di nascita e di morte, che quelle di carattere "strutturale", cioè le relazioni genitori-figli e marito-moglie. Per la sua descrizione formale si veda www.gedcom.org.

[6] In realtà, almeno per quanto riguarda i registri parrocchiali di Marciana su cui sta lavorando l'autore, non sempre le informazioni sono così complete: l'indicazione della madre degli sposi, per esempio, comincia a essere frequente solo dopo la metà del XVIII secolo;

nome madre sposo, cognome madre sposo, provenienza sposo, nome sposa, padre sposa, cognome sposa, nome madre sposa, cognome madre sposa, provenienza sposa.

Una registrazione come questa fa sì che vengano aggiunte al database sei nuove persone. Ma sono veramente nuove persone? Supponiamo che lo sposo o la sposa abbiano dei fratelli che si sono sposati qualche tempo prima; in questo caso i genitori saranno già presenti in archivio e, con la nuova registrazione di matrimonio, verrà creato un duplicato dei genitori. Inoltre, uno o entrambi gli sposi potrebbero essere alle seconde nozze e quindi anche i loro nomi si troverebbero già nel database; lo stesso accadrebbe se fosse stato già registrato il loro battesimo[7]. Dunque, via via che vengono caricate nel database le registrazioni di battesimi, matrimoni e defunti, inevitabilmente si creano numerosi duplicati della stessa persona. A volte, soprattutto nei casi di persone che hanno molti figli e nipoti, nei registri parrocchiali potremmo trovare decine di citazioni, ognuna delle quali corrisponde a una nuova copia della stessa persona!

Come affrontare questo problema? Una strategia consiste nell'evitare fin dall'inizio di inserire doppioni, costruendo in tal modo un database sostanzialmente "pulito" e pronto per essere usato nei diversi contesti disciplinari della demografia storica. Di solito questo modo di procedere richiede che sia un operatore umano a valutare di volta in volta se il nominativo da inserire è già presente nel database o se si tratti di una nuova persona; infatti, un programma avrebbe grande difficoltà a raggiungere questo obiettivo, soprattutto se i dati disponibili sono frammentari e imprecisi. Questa è la strategia che abbiamo definito come "interattiva", e sulla quale non ci soffermeremo.

Vedremo nel seguito che l'approccio del software Sage è invece principalmente di tipo automatico, anche se lascia aperta la possibilità di un intervento diretto da parte dell'operatore umano. Occorre sottolineare che questo intervento diretto è indispensabile per risolvere le situazioni di ambiguità, o i veri e propri errori, che possono manifestarsi. In pratica il programma viene mandato in esecuzione e lasciato lavorare per tutto il tempo necessario; al termine si esaminano, con opportuni strumenti software, i risultati ottenuti, si effettuano le eventuali correzioni dei dati, e si procede a un nuovo *run*.

Dunque, più che di una strategia totalmente automatica, nel senso indicato da Schofield, o di una interattiva, è più appropriato parlare di una stra-

persino i cognomi degli sposi cominciano a essere citati costantemente solo dopo la metà del XVII secolo. In effetti in epoca più antica la popolazione di Marciana era piuttosto ridotta, appena poche centinaia di persone e, a quanto pare, il sacerdote si limitava a riportare il minimo di informazioni necessarie affinché i membri della comunità potessero identificare gli sposi.

[7] Nello studio di una piccola comunità come Marciana, quasi sempre è possibile ritrovare la registrazione del battesimo degli sposi.

tegia "iterativa", nel senso che il risultato finale di un database (ragionevolmente) "pulito" può essere ottenuto solo attraverso un certo numero, auspicabilmente finito, di iterazioni.

3. *Il modulo costruttore*. Nella strategia di Sage, ogni ciclo di elaborazione inizia con l'esecuzione di un modulo, che abbiamo definito "costruttore", il quale ogni volta crea *ex novo* il database Gedcom a partire dal foglio elettronico. Come si è già accennato, gli interventi manuali per la correzione dei dati sono indispensabili e, anche se in teoria potrebbero essere effettuati direttamente sul database Gedcom, risulta molto più pratico e meno soggetto a errori intervenire su un foglio elettronico. Occorre quindi tradurre le informazioni da una forma all'altra, e questo è proprio il compito del modulo "costruttore", chiamato così appunto perché costruisce il database. A partire dai dati di una singola registrazione documentale, ossia da una riga del foglio elettronico, esso crea e inserisce nel database le persone coinvolte e le rispettive relazioni famigliari, corredandole di tutte le informazioni che è possibile dedurre direttamente o indirettamente dalla registrazione stessa.

Detto questo, è opportuno osservare che questa modalità di lavoro, basata sul foglio elettronico e il costruttore, è solo una scelta personale dell'autore; infatti, il programma unificatore in teoria potrebbe lavorare su qualsiasi file Gedcom che gli venga fornito in input.

In un database Gedcom ogni persona è caratterizzata da un certo insieme di attributi e "creare" una persona significa semplicemente assegnare un valore appropriato a tali attributi. Essi sono di due tipi: quelli essenziali e quelli aggiuntivi. Attributi essenziali sono: il codice, che è un numero progressivo univoco assegnato dal programma, il nome, il sesso, la data di nascita (effettiva o stimata) e il riferimento al documento in cui la persona è citata. Attributi aggiuntivi, così detti perché non sempre presenti, sono: cognome, luogo di provenienza (se diverso da Marciana), data di morte (effettiva o stimata), famiglia di origine e famiglia (o famiglie) discendenti.

Analogamente, le famiglie hanno sia attributi essenziali che aggiuntivi; quelli essenziali sono il codice progressivo, assegnato automaticamente, e il marito. Il marito viene sempre considerato esistente, indipendentemente dal fatto che all'epoca cui si riferisce la registrazione egli sia vivo o morto. La moglie invece è considerata tra gli attributi aggiuntivi in quanto possono esistere sia famiglie complete, nelle quali il suo nome è conosciuto, sia famiglie incomplete, in cui si conosce solo il nome del padre[8]. Quando il padre non è conosciuto, come nei battesimi di neonati abbandonati, o dei pochissimi

[8] Le famiglie incomplete, che alla fine del processo di unificazione costituiscono circa il 20% del totale, vengono generate automaticamente ogni volta che in una registrazione

casi di figli di donne non sposate, viene creata una famiglia in cui al padre è assegnato il nome in codice "incerto"[9].

Anche la data di matrimonio e il riferimento al relativo documento sono considerati attributi aggiuntivi, in quanto non sempre sono noti. Nei casi in cui manchi la registrazione documentale, l'anno di matrimonio viene stimato, ma solo se la famiglia è completa. Analogamente i figli vengono considerati attributi aggiuntivi, dal momento che possono essere presenti o meno. Nei casi in cui una persona contragga matrimoni successivi, ogni volta verrà costituita una nuova famiglia, senza che quella precedente sia cancellata. Naturalmente ciò presuppone che il coniuge precedente sia morto. Se non viene identificato il relativo documento nel registro dei defunti, il programma stesso provvede a stabilirne la morte in data precedente alla nuova unione. Quindi una persona può generare figli in famiglie diverse e successive, ma i figli nati in ciascuna di esse manterranno la loro appartenenza a quella specifica in cui sono nati.

Vediamo adesso di chiarire meglio il processo di costruzione del database Gedcom mediante un esempio tratto dal registro dei defunti[10]. Nella figura 1 sono schematizzate tutte le fasi operative, dalla decodifica del documento all'inserimento nel database delle persone coinvolte.

Nella figura 1a) le informazioni rilevanti vengono ricavate manualmente dal documento originale e inserite nel foglio elettronico, in modo che a ogni registrazione corrisponda una riga. Successivamente il programma costruttore, elaborando questa specifica riga del foglio elettronico, produce lo spezzone di file Gedcom riportato nella figura 1b). Lo schema in fig. 1c), introdotto solo a scopo esplicativo, presenta in forma grafica il contenuto di questo spezzone, cioè la struttura familiare deducibile dal documento[11].

Vediamo adesso più in dettaglio come il costruttore ha elaborato le informazioni contenute nella registrazione. Per prima cosa ha creato la persona Anna Lupi e le informazioni che la riguardano sono contenute nelle righe da 1 a 11 del file riprodotto in fig. 1b). Nella prima riga troviamo il numero 42228 che è il codice progressivo assegnato automaticamente alla persona. Considerato che il documento è tratto dal registro defunti, l'unica data che possiamo considerare certa è quella di sepoltura, che è riportata nella riga 7;

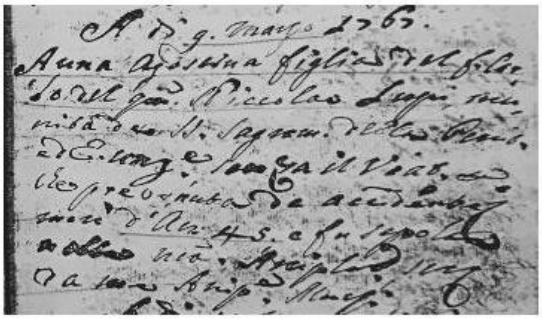
compare un patronimico: per es. Antonio di Giovanni, senza altre indicazioni. Questi casi riguardano principalmente le epoche più antiche e i forestieri.

[9] Quando il programma unificatore incontra il nome "incerto" evita di cercare un collegamento con altre persone.

[10] Archivio parrocchiale, Chiesa di santa Caterina, Marciana (d'ora in poi Apscm), *Defunti*, vol. II, p. 215, n. 5.

[11] Lo schema non è prodotto dal programma costruttore, ma da un'apposita interfaccia di interrogazione grafica del database, che utilizza il linguaggio Dot sviluppato nell'ambito del progetto *open source* Graphviz.

Fig. 1. Esempio di funzionamento del costruttore dei defunti



A di 9 marzo 1767
Anna Agostina figlia del fu Carlo di Niccolao Lupi munita dei SS Sacramenti della Penitenza ed Estrema Unzione senza il Viatico perché prevenuta da accidente morì d'anni 45 e fu sepolta nella nostra [chiesa] Arcipretale ut super da me Arciprete Murzi

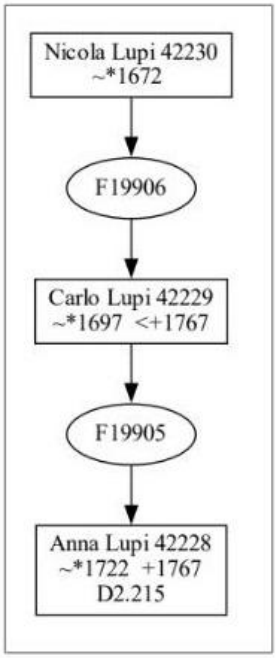
a)

	A	B	C	D	E	F	G	H	I
1	**RIF**	**DATA**	**NOME**	**S**	**FU**	**PADRE**	**NONNO**	**COGNOME**	**ETA**
2	D2.215	09/03/1767	Anna Agostina	F	fu	Carlo	Niccolao	Lupi	45

b)

```
0 @I42228@ INDI
1 NAME Anna \Lupi\
1 SEX F
1 BIRT
2 DATE DEC 1722
1 DEAT
2 DATE 09/03/1767
2 TYPE *D2.215* morte di Anna Lupi (45)
1 CITA
2 NOTE 1767*defunto*D2.215
1 FAMC @F19905@
0 @I42229@ INDI
1 NAME Carlo \Lupi\
1 SEX M
1 BIRT
2 DATE CAL 1697
1 DEAT
2 DATE BEF 09/03/1767
1 CITA
2 NOTE 1767*padre def*D2.215
1 FAMC @F19906@
1 FAMS @F19905@
0 @F19905@ FAM
1 HUSB @I42229@
1 CHIL @I42228@
0 @I42230@ INDI
1 NAME Nicola \Lupi\
1 SEX M
1 BIRT
2 DATE CAL 1672
1 CITA
2 NOTE 1767*nonno pat def*D2.215
1 FAMS @F19906@
0 @F19906@ FAM
1 HUSB @I42230@
1 CHIL @I42229@
```

c)



nella riga 10 troviamo il riferimento del documento. La data di nascita della persona non è conosciuta, ma dall'età di morte dichiarata (45 anni) si risale all'anno di nascita, riportato nella riga 5; il codice "dec 12" indica che si tratta di una data «dichiarata», cioè una stima abbastanza credibile[12].

Per quanto riguarda il nome Anna, notiamo che nella registrazione era invece riportato Anna Agostina; analogamente il nome del nonno era Niccolao, mentre nel file (riga 27) troviamo Nicola. Ci sono queste differenze perché uno dei compiti fondamentali del costruttore è la standardizzazione dei nomi e cognomi. A questo proposito il metodo utilizzato dal sistema Sage è molto semplice: si basa su un file di sinonimi, dove è scritto che Niccolao diventa Nicola, che Anna Agostina diventa Anna, che il cognome Pavolini diventa Paolini e così via.

Questa soluzione così elementare non è però esente da inconvenienti. Per esempio, nel caso dei nomi plurimi, che cominciano a diffondersi dopo la metà del XVIII secolo, succede che al battezzato vengano dati anche 3 nomi e poi, nel corso della vita, non viene usato il primo ma il secondo o anche il terzo. Inoltre, nel caso di nomi doppi come Gio. Giuseppe, capita che in alcune registrazioni siano riportanti entrambi e in altre solo il primo o solo il secondo. Queste situazioni danno luogo inevitabilmente a condizioni di errore che possono essere risolte solo con un intervento manuale. In generale il problema della standardizzazione dei nomi e cognomi rimane comunque una questione ancora aperta[13].

Il costruttore si occupa anche di un altro aspetto importante che è quello della documentazione, ossia l'inserimento automatico degli estremi del documento in cui una persona è citata, insieme al ruolo ricoperto nell'evento oggetto della registrazione. Queste informazioni, contenute nel blocco "1 cita" e nei suoi sotto elementi "2 note"[14], sono utili al ricercatore per correggere eventuali problemi nel database, in quanto consentono di identificare immediatamente il documento originale. Tornando all'esempio, vediamo quindi che Anna Lupi è citata nell'anno 1767 in qualità di «defunto» nella registrazione D2.215[15]; analogamente Carlo Lupi è citato come «padre def»

[12] A questo proposito c'è da dire che nei registri di Marciana l'età del defunto comincia a essere riportata con frequenza solo verso la fine del XVII secolo; inoltre, molto spesso tale indicazione è poco accurata. Confrontando le età dichiarate a quelle effettive ricavate dai battesimi (nei casi in cui il documento viene identificato), si riscontra quasi sempre una differenza di alcuni anni, generalmente in eccesso rispetto al valore reale.

[13] Si veda per esempio R. Abramitzky, R. Mill, S. Pérez, *Linking Individuals Across Historical Sources: A Fully Automated Approach*, in «Historical Methods», 53, 2020, 2, pp. 94-111.

[14] 1 Note e 2 Cita non fanno parte dello standard Gedcom e sono state aggiunte dall'autore.

[15] D2.215 indica registro defunti già citato; con questo riferimento, in caso di dubbio è sempre possibile controllare il testo originale del documento.

Fig. 2. Ordinamento del foglio Excel per data (a) e per sequenza familiare (b)

a)

1	RIF	DATA	NOME	S	PADRE	NONNO P.	COG. P.	MADRE	NONNO M.	COG. M.
2	B5.4.1	04/09/1784	Lorenzo	M	Lorenzo	Gio	Palillo	Giovanna Dor	Francesco	Provenzali
3	B5.4.2	05/09/1784	Francesco	M	Giuseppe	Pietro	Giudicelli	Maria	Francesco	Costa
4	B5.4.3	08/09/1784	Oliva	F	Andrea	Domenico	Murzi	Felicia	Francesco	Pierangeli
5	B5.4.4	15/09/1784	Gio Antonio	M	Gio	Francesco	Paolini	Simona	Liborio	Retali
6	B5.4.5	22/09/1784	Nicola	M	Giuseppe	Aurelio	Galanti	Giovanna	Giuseppe	Pavolini
7	B5.5.1	30/09/1784	Fortunata	F	Bartolomeo	Fortunato	Marchiani	Agata	Giacomo	Carnevali
8	B5.5.2	10/10/1784	Cerbone	M	Domenico	Luca	Lupi	Giovanna	Gio Domenico	Arnaldi
9	B5.5.3	28/10/1784	Santa	F	Nicola	Domenico	Braschi	Bartolomea	Domenico	Sardi
10	B5.5.4	02/11/1784	Santa	F	Natale	Paolo	Cianchini	Giuseppa	Marco	Bianchi
11	B5.6.1	09/11/1784	Pellegrino	M	Nicola	Domenico	Berti	Caterina	Lazzaro	Cauci
12	B5.6.2	28/11/1784	Frediano	M	Gio	Carlo	Pierangeli	Giovanna	Gio	Pavoni
13	B5.6.3	28/11/1784	Francesco	M	Valentino	Pierulivo	Pierulivo	Margherita	Paolo	Garbati
14	B5.6.4	29/11/1784	Caterina	F	Giuseppe	Gio	Anselmi	Barbara	Francesco	Pavolini
15	B5.7.1	30/11/1784	Tommaso	M	Natale	Tommaso	Lupi	Chiara	Silvestro	Poggioli
16	B5.7.2	17/01/1785	Caterina	F	Gio	Domenico	Braschi	Gerolama	Antonio	Pieruzzini
17	B5.7.3	20/01/1785	Sebastiana	F	Gio	Antonio	Testa	Antonia		Pieruzzini
18	B5.8.1	23/01/1785	Giuseppe	M	Amedeo		Pisani	Antonia	Martino	Murzi
19	B5.8.2	24/01/1785	Sebastiano	M	Gio	Nicola	Murzi	Vittoria	Gio	Sardi

b)

1	RIF	DATA	NOME	S	PADRE	NONNO P.	COG. P.	MADRE	NONNO M.	COG. M.
2	B5.252.4	25/06/1797	Giuseppe	M	Agabito	Giuseppe	Peria	Giuseppa	Filippo	Sardi
3	B5.288.3	04/08/1799	Gio Giuseppe	M	Agabito	Giuseppe	Peria	Giuseppa	Filippo	Sardi
4	B5.337v.1	04/03/1804	Gio	M	Agabito	Giuseppe	Peria	Giuseppa	Filippo	Sardi
5	B6.119.1	13/12/1807	Giuseppe	M	Agabito	Vincenzo	Anselmi	Maria	Gio	Ciangherott
6	B8.p024.1	18/03/1811	Elisabetta	F	Agabito	Vincenzo	Anselmi	Maria	Gio	Ciangherott
7	B8.p050.2	31/01/1813	Elisabetta	F	Agabito	Vincenzo	Anselmi	Maria	Gio	Ciangherott
8	B8.p124.1	16/06/1816	Francesco	M	Agabito	Vincenzo	Anselmi	Maria	Gio	Ciangherott
9	B6.169.1	24/07/1809	Giuseppe	M	Agostino	Emiliano	Berti	Benedetta	Gio Lorenzo	Piola
10	B8.p017.4	14/10/1810	Giuseppe	M	Agostino	Emiliano	Berti	Benedetta	Gio Lorenzo	Piola
11	B8.p137.4	07/01/1817	Felicia	F	Agostino	Emiliano	Berti	Benedetta	Gio Lorenzo	Piola
12	B6.70.2	31/08/1806	Marianna	F	Agostino	Michelangelo	Sardi	Camilla	Antonio	Berti
13	B6.139.2	25/07/1808	Angela	F	Agostino	Michelangelo	Sardi	Camilla	Antonio	Berti
14	B8.p033.4	21/02/1810	Caterina	F	Agostino	Michelangelo	Sardi	Camilla	Antonio	Berti
15	B6.86.1	23/02/1807	Francesco	M	Agostino	Giacomo	Giudice	Caterina		Fantini
16	B5.329v.3	31/07/1803	Giacomo	M	Agostino	Giacomo	Giudici	Caterina	Francesco	Fantini
17	B6.9.2	04/12/1804	Maria Domeni	F	Agostino	Giacomo	Giudici	Caterina		Fantini
18	B5.158.2	20/02/1793	Maria	F	Agostino	Arcangelo	Fossi	Francesca	Pietro	Cauci
19	B5.213.1	24/08/1795	Arcangelo	M	Agostino	Arcangelo	Fossi	Francesca	Pietro	Cauci

e Nicola Lupi come «nonno pat def», sempre con indicazione dell'anno e del documento originale.

Inoltre, il costruttore spesso può rendere meno gravoso il successivo lavoro del modulo unificatore; infatti, in molti casi è in grado di procedere direttamente alla formazione delle famiglie. Ciò avviene in particolare con i battesimi[16]. Per comprendere la strategia utilizzata, occorre soffermarci sull'importanza del modo in cui viene ordinato l'archivio. In fase di inserimento manuale delle registrazioni, l'ordine segue quello con cui si presentano le registrazioni stesse, ossia quello per numero di pagina del registro. Tale ordinamento, che quasi sempre coincide con quello temporale, non è quello otti-

16 In teoria si potrebbe procedere nello stesso modo anche per defunti e matrimoni, ma in pratica risulta molto più difficile soprattutto per le epoche più antiche, in quanto le registrazioni sono molto scarne e spesso sono citati solo i diretti interessati, senza riferimento ai genitori e tanto meno ai nonni.

male per i nostri scopi. Infatti, se si desidera ricostruire le famiglie, occorre identificare tutti i figli nati in anni diversi da una certa coppia. Ciò può essere ottenuto con facilità riordinando in maniera opportuna l'archivio. Il criterio utilizzato è riportato di seguito, con gli attributi elencati in ordine gerarchico: 1) nome del padre, 2) nome della madre, 3) cognome padre, 4) cognome madre, 5) nome nonno paterno, 6) nome nonno materno, 7) anno. Questa sequenza è in grado di identificare, quasi senza alcun intervento manuale[17], i figli nati da una certa coppia di genitori ossia, in definitiva, ricostruire automaticamente le famiglie. L'effetto di questo criterio di ordinamento si può osservare in fig. 2.

Il metodo usato dal costruttore per identificare una famiglia è abbastanza semplice. Esaminando l'elenco dei battesimi trova il punto in cui cambia il nome del padre o della madre, o quello dei nonni, o il loro cognome e, in base a tali "rotture" della sequenza, definisce i confini della famiglia. Per esempio, osservando la figura 2b), notiamo che la prima famiglia, quella che ha per padre Agabito Peria e per madre Giuseppa Sardi, è composta da 3 figli: Giuseppe, Gio. Giuseppe e Gio. Analogamente la seconda, che ha per padre Agabito Anselmi e madre Maria Ciangherotti, è composta dai figli Giuseppe, Elisabetta, Elisabetta e Francesco. Il risultato finale è schematizzato in fig. 3.

Nel costruire una famiglia, il programma cura anche altri dettagli essenziali per la coerenza del database. Per esempio, notiamo che nella seconda famiglia troviamo due volte il nome Elisabetta; siccome in una famiglia non possono esserci due figli con lo stesso nome, si deduce che la Elisabetta nata nel 1811 deve essere morta prima del 31 gennaio 1813, data di nascita della seconda Elisabetta. Il programma provvede quindi automaticamente a fissare per la prima Elisabetta una data di morte precedente al 31 gennaio 1813.

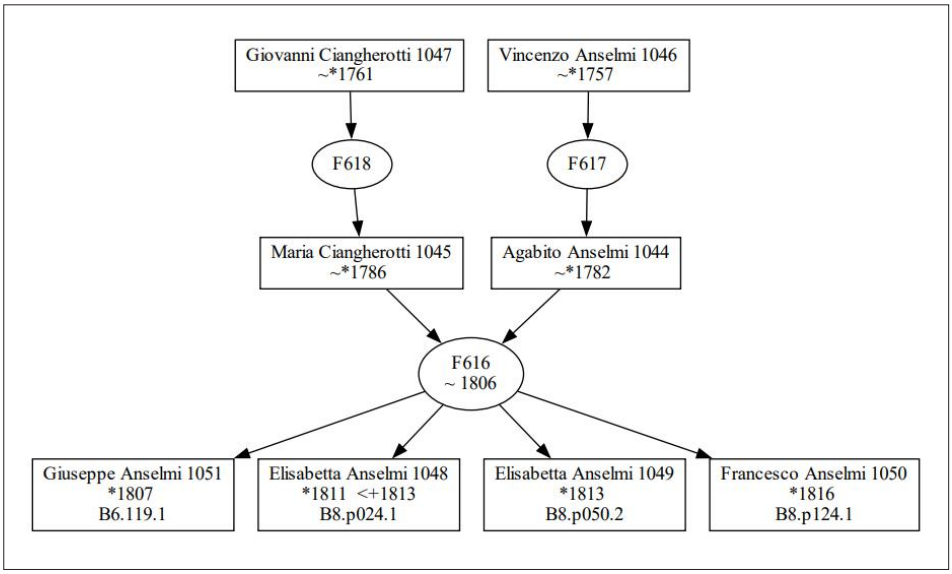
Osserviamo che nello schema le uniche date certe sono quelle dei battesimi[18]; a partire da esse il programma ha stimato la data del matrimonio, che è assunta un anno prima della nascita del primo figlio, l'anno di nascita del padre, 24 anni prima del matrimonio, quello della madre, 20 anni prima, e quelle dei nonni paterni, rispettivamente 25 anni prima dei due genitori.

Qual è l'utilità effettiva di questa procedura di costruzione delle famiglie? Dall'esempio considerato si vede che ogni registrazione, oltre al battezzato, riporta quattro persone, ossia i genitori e i nonni paterni. Se le quattro registrazioni battesimali venissero considerate indipendentemente una dall'altra, cioè come si presentano in ordine di tempo, il costruttore dovrebbe inserire nel database 4 volte la stessa famiglia, ossia 20 persone, mentre procedendo

[17] L'intervento manuale è indispensabile solo nei casi in cui si incontrano genitori con nomi molto comuni o, come avviene spesso nelle registrazioni più antiche, mancano i cognomi e talvolta anche i nomi dei nonni.

[18] La presenza del carattere ~ a precedere indica una data stimata.

Fig. 3. Esempio di famiglia generata dal costruttore dei battesimi



come descritto sopra le persone da inserire diventano solo 8. Il risparmio è ancora più consistente nel caso di famiglie che hanno avuto molti figli oppure quando nelle registrazioni vengono citate anche le nonne materne o i bisnonni. Siccome il tempo di elaborazione del programma di unificazione cresce con il numero di persone secondo una legge grosso modo esponenziale, il risparmio che si ottiene con questo lavoro preliminare è notevole.

È importante notare che durante la fase di unificazione, il programma tenterà poi di arrivare a una identificazione precisa di tutte le informazioni che in questo primo momento sono state lasciate in sospeso. In particolare, cercherà di trovare la registrazione del matrimonio, nonché il battesimo dei genitori e dei nonni, via via risalendo a ricostruire, almeno in teoria, tutta la rete di parentele.

4. *Il programma unificatore*. Una volta che i programmi costruttori hanno creato il database Gedcom, inizia la fase di unificazione. Come vedremo essa procede in ordine di nome, quindi per prima cosa il programma crea un file sequenziale contenente una "vista inversa"[19] del database ordinata per sesso,

[19] Per "vista inversa" di un database si intende un diverso modo di vedere gli stessi dati contenuti nel database, al fine di rendere più agevole una certa operazione.

Fig. 4. Frammento del database ordinato per sesso, nome e anno di nascita (a); legenda (b)

a)

```
1  M*Vittorio*
2  M*Vittorio*1537**Luca**28*******^1577*stimatore*Doti p6
3  M*Vittorio*1540****6080****Cerbone***^1590*nonno pat batt*B1. 21. 103^1596*nonno pat batt*B1. 31. 1
4  M*Vittorio*1541****387****Cerbone***^1591*padre sposo*Doti p66v
5  M*Vittorio*1581**Trolio**26082**Lorenza*****^1606*sposo*M1. 25. 101
6  M*Vittorio*1582**Domenico**25691**Lorenza**Bartolomea,Giovanni***^1607*padre batt*B1. 52. 368^1614*
7  M*Vittorio*1585*Sardi***43262**Giacoma*****^1646*marito defunta*D1. 34_586
8  M*Vittorio*1587**Trolio**42290**Lorenza*****^1615*marito defunta*D1.  7_132
9  M*Vittorio*1590**Cerbone*Eleonora*6082*****Vittorio***^1590*batt *B1. 21. 103
10 M*Vittorio*1591*Sardi***11605****Caterina***^1637*nonno mat batt*B1.186.1086
11 M*Vittorio*1591*Sardi*Cerbone**26249**Giacoma*****^1616*sposo*M1. 36. 145
12 M*Vittorio*1592*Sardi*Cerbone**25673**Giacoma**Caterina,Rosa,Giovanna,Filippa,Gerolama***^1617*padr
13 M*Vittorio*1592*Sardi*Cerbone**26423**Giacoma**Caterina***^1637*padre sposa*M1. 50. 190
14 M*Vittorio*1593**Cerbone**788**Giacoma*****^1618*sposo*Doti p139v
15 M*Vittorio*1593*Sardi*Cerbone**26452**Giacoma**Rosa***^1638*padre sposa*M1. 51. 197
16 M*Vittorio*1596*Sardi***5020****Rosa***^1642*nonno mat batt*B1.216.  78
17 M*Vittorio*1597*Sardi*Cerbone**26553**Giacoma**Filippa***^1642*padre sposa*M2.  5. 421
18 M*Vittorio*1602*Sardi*Cerbone**26623**Giacoma**Filippa***^1647*padre sposa*M2.  9. 440
19 M*Vittorio*1607*Gentili***44357**Mattea*****^1674*marito defunta*D1. 62_1121
20 M*Vittorio*1610*Gentili***25708**Mattea**Monica,Nicola,Guglielmo,Apollonia,Domenico,Cecilia***^1635
21 M*Vittorio*1613*Gentili*Nicola*Camilla*20927*****Matteo***^1613*batt *B1. 78. 530
22 M*Vittorio*1620*Gentili***19175****Lucrezia***^1666*nonno mat batt*B1.274. 628
23 M*Vittorio*1621*Ferrini*Francesco**43623***18/09/1649****^1649*defunto*D1. 44_763
24 M*Vittorio*1621*Ferrini*Francesco*Innocenzia*9911*****Giovanni***^1621*batt *B1.122. 732
25 M*Vittorio*1625*Gentili***27032****Lucrezia***^1670*padre sposa*M1. 66. 257
26 M*Vittorio*1628*Gentili***27206**Mattea**Cecilia***^1678*padre sposa*M1. 77. 307
27 M*Vittorio*1630*Gentili***27162****Camilla***^1675*padre sposa*M1. 74. 294
28 M*Vittorio*1631*Gentili***44218***22/03/1671*****^1671*defunto*D1. 59_1054
29 M*Vittorio*1633*Gentili***27864****Domenico***^1708*nonno sposo*M2. 72. 184
30 M*Vittorio*1636*Gentili***27788****Domenico***^1706*nonno sposa*M2. 67. 167
31 M*Vittorio*1642*Palmieri*Bartolomeo**43103*S.margherita**11/05/1643*****^1643*defunto*D1. 29_506
32 M*Vittorio*1642*Palmieri*Bartolomeo*Rosa*5023*****Agostino***^1642*batt *B1.216.  78
33 M*Vittorio*1644*Baccelli*Filippo**45196***23/10/1699*****^1699*defunto*D2. 30_1541 (55)
34 M*Vittorio*1652*Ferrini*Lorenzo**45027***26/10/1695*****^1695*defunto*D2. 18_1459 (43)
```

b)

1	2	3	4	5	6	7	8	9	10	11	12	13	14
sex	nome	anno	cognome	padre	madre	progressivo	luogo	nome coniuge	data morte	elenco figli	nonno	cog.coniuge	elenco citazioni

nome e anno di nascita. Vediamo in fig. 4 un frammento di questo file relativo al nome Vittorio, insieme alla sua struttura logica indicata nella legenda[20].

Senza indagare tutti i dettagli, possiamo osservare che le prime occorrenze del nome Vittorio mancano del cognome e al suo posto troviamo spesso il nome del padre[21]. Notiamo anche, nella parte terminale della riga, l'elenco delle citazioni con il riferimento al documento e il ruolo della persona; quasi tutti hanno una sola citazione, perché il costruttore ha inserito nel database una nuova occorrenza per ognuna delle persone citate nell'evento. Unica eccezione sono i genitori che hanno avuto più figli, perché in questo caso, come già esposto, il programma ha provveduto a costruire la famiglia discendente. Per esempio, nella riga 6 della fig. 4a) vediamo che un certo Vittorio di Domenico, nato intorno al 1582, con moglie Lorenza, ha avuto i figli Bartolomea nel 1607 e Giovanni nel 1614.

L'occhio attento noterà che alcune occorrenze sono riferite quasi certa-

[20] I campi sono separati dal carattere asterisco. Se il campo è vuoto si vedono due asterischi vicini.

[21] In effetti a Marciana i cognomi cominciano a diffondersi nella prima metà del XVII secolo e per le registrazioni precedenti è più frequente l'uso del patronimico.

mente alla stessa persona; per esempio, alla riga 11 troviamo un Vittorio Sardi, nato intorno al 1591, che nel 1616 si sposa con una certa Giacoma. Nella riga successiva, di nuovo troviamo un Vittorio Sardi, nato circa nel 1592, con moglie Giacoma, che dal 1617 in poi ha avuto le figlie Caterina, Rosa, Giovanna, Filippa e Gerolama. E qui dobbiamo affrontare un nodo essenziale: possiamo concludere che si tratta della stessa persona? Per rispondere conviene esaminare gli elementi a nostra disposizione.

i) Il nome Vittorio. La domanda da porsi è: si tratta di un nome raro? Se la risposta fosse affermativa, cioè se questo nome comparisse solo poche volte nelle registrazioni, allora già questo costituirebbe una coincidenza importante. Ma come si può stabilire se un nome è raro o no? La risposta evidentemente si basa su una preliminare analisi della frequenza dei nomi e dei cognomi. Da essa risulta che il nome maschile più comune è Giovanni, con 3553 occorrenze[22]; in maniera abbastanza arbitraria si è scelto di considerare un nome "non comune" se la sua frequenza è 20 volte inferiore (ovvero 177 occorrenze) e "raro" se è 50 volte inferiore (71 occorrenze). Il nome Vittorio compare 116 volte, quindi è da considerare "non comune", dunque possiamo considerare quella del nome una coincidenza abbastanza significativa. Come vedremo in seguito, il programma di unificazione attribuisce un punteggio a tutta una serie di parametri e tra questi rientra anche la frequenza del nome.

ii) Il cognome Sardi. Questo cognome è uno dei più comuni a Marciana e quindi la coincidenza non è da considerare significativa.

iii) L'anno di nascita stimato è quasi lo stesso. Questo dovrebbe escludere che si tratti di un nonno o di un padre con lo stesso nome[23], pur tenendo presente l'ampia incertezza nella stima dell'anno di nascita; tuttavia, non si può escludere che si tratti di un cugino omonimo nato nello stesso periodo. In conclusione, questa coincidenza è significativa ma non particolarmente.

iv) Entrambi hanno una moglie di nome Giacoma. Questa coincidenza ha un peso decisivo; infatti, in una popolazione relativamente piccola[24] come quella di Marciana, la coincidenza di nome, cognome e nome del coniuge è estremamente rara e si verifica solo con nomi molto comuni.

Possiamo quindi concludere che vi è una probabilità molto alta che si trat-

22 Su un totale di 57.752 persone attualmente presenti in archivio, considerando insieme maschi e femmine.

23 La consuetudine di assegnare al primogenito il nome del nonno, più spesso quello paterno, era molto comune a Marciana, tanto è vero che in alcune famiglie è possibile riconoscere per diversi secoli il nome dell'avo; meno frequente è invece l'attribuzione dello stesso nome del padre.

24 Secondo le stime dell'autore nella prima metà del XVII secolo la popolazione era di circa 550 persone (C. Anselmi, *La crisi di mortalità del 1647 a Marciana*, in «Proposte e ricerche», n. 89, 2022, pp. 151-169).

ti della stessa persona, per cui considereremo valida l'unificazione. Qui occorre, però, affrontare una questione delicata. Le scelte di questo tipo sono fatte sempre e comunque su base probabilistica e, per di più, senza poter calcolare esattamente la probabilità sottostante. Per essere più precisi, dal punto di vista matematico si tratta di una probabilità "soggettiva"[25]. Dunque, lo storico deve accontentarsi di prendere una decisione di carattere probabilistico, basandosi sulla quantità e qualità delle coincidenze. Come vedremo, anche il programma unificatore si comporta in maniera sostanzialmente simile.

Osservando la fig. 4 si nota che le occorrenze nel nome Vittorio sono ordinate per anno di nascita crescente. Ciò è necessario per il modo di operare del programma di unificazione. In effetti esso procede ciclicamente, esaminando una persona alla volta e cercando nell'elenco altre persone con cui unificarla. In altre parole, prende il primo nome della lista, che chiameremo "soggetto", e guarda se tra quelli che seguono in ordine di tempo, che chiameremo "candidati", ne trova qualcuno adatto. Poiché l'incertezza nella stima dell'anno di nascita può essere notevole, il programma si spinge a considerare come possibili candidati quelli nati fino a 50 anni dopo[26].

In questo modo viene creato una prima lista grezza di candidati che sono compatibili almeno dal punto di vista temporale. In questo processo vengono automaticamente scartati tutti quelli che, per ovvi motivi, risultano incompatibili con il soggetto. In tabella 1 è riportato un elenco dei principali motivi di esclusione.

Tab. 1. Principali criteri di esclusione

soggetto	*candidato*
ha il cognome	ha un cognome diverso
ha il padre	ha un padre diverso
ha la madre	ha una madre diversa
ha il nonno paterno	ha un nonno paterno diverso
è forestiero	è forestiero con provenienza diversa
ha data di nascita certa	ha data di nascita certa e diversa
ha data di morte certa	ha data di morte certa e diversa
ha un figlio con lo stesso nome del candidato	è il figlio del soggetto
ha data di morte	è nato dopo la morte del soggetto
ha data di morte	ha avuto figli dopo la morte del soggetto

[25] È lo stesso tipo di probabilità che riguarda, per esempio, gli eventi sportivi. Se si organizza una partita di beneficenza tra la squadra di calcio che ha vinto il campionato di serie A e la squadra amatoriale dei cantanti, nessuno può calcolare esattamente la probabilità che vinca la squadra di serie A, ma quasi tutti sono disposti a scommettere che vincerà lei.

[26] Tra i motivi che rendono necessario considerare un intervallo temporale così ampio c'è anche il fatto che a volte si riscontra una forte differenza di età tra i coniugi, che può raggiungere i 20 o 30 anni, soprattutto nelle seconde nozze.

ha data di morte	si è sposato dopo la morte del soggetto

In realtà questi semplici criteri non sono sufficienti a garantire che a seguito dell'unificazione non si creino delle incoerenze nel database. Per esempio, può succedere che il padre e il nonno siano compatibili, ma non il bisnonno o un altro antenato, o un parente qualsiasi. Vedremo più avanti quali meccanismi sono previsti per fare fronte a queste evenienze.

Dunque, dopo questa prima fase di selezione, il programma ha a disposizione una lista più o meno lunga di candidati e il problema diventa ora quello di scegliere quello che ha la maggior probabilità di essere quello giusto. La strategia utilizzata da Sage è quella di assegnare un punteggio a ogni candidato in base a una serie di criteri, ognuno dei quali può contribuire positivamente o negativamente. Alla fine, verrà scelto quello con il punteggio più alto, purché questo punteggio sia superiore a una certa soglia minima prestabilita. Nella tabella 2 è riportato un elenco dei principali criteri usati per l'attribuzione del punteggio.

Tab. 2. Principali criteri per l'attribuzione del punteggio

criterio	*punti*
nome non comune	2
nome raro	5
cognome non comune	2
cognome raro	5
soggetto e candidato hanno entrambi il cognome coincidente	1
solo uno dei due ha il cognome (per non favorire indebita propagazione del cognome)	-1
soggetto e candidato hanno entrambi cognome e padre coincidenti	1
soggetto e candidato hanno entrambi padre e nonno coincidenti	1
entrambi sono forestieri provenienti dallo stesso luogo	4
soggetto e candidato sono nati nello stesso anno	1
differenza di età >40 anni	-25
differenza di età >30 anni	-6
differenza di età >25 anni	-3
differenza di età >15 anni	-2
differenza di età >10 anni	-1
presenti figli con lo stesso nome del padre	1
presenti figli con lo stesso nome della madre	1
presenti figli con lo stesso nome del nonno o della nonna paterna	1
presenti figli con lo stesso nome del nonno o della nonna materna	1
soggetto e candidato hanno figli con lo stesso nome	1
uno ha il coniuge e l'altro no	-1
soggetto e candidato hanno un coniuge con lo stesso nome	2

Una volta che è stato identificato il candidato migliore, inizia il processo di unificazione vero e proprio, di cui parleremo tra breve, dopo di che il candidato prescelto verrà cancellato dalla lista. In tal modo la lista stessa si accorcia progressivamente e alla fine conterrà solo il soggetto, oppure altre persone che però non sono unificabili con esso. Quando il programma ha terminato le possibili unificazioni che riguardano un certo soggetto, passa a considerare la persona successiva con lo stesso nome o, se non ce ne sono altre, al nome successivo, procedendo nello stesso modo finché non sono state esaminate tutte le persone presenti nell'archivio.

Vediamo ora in maggior dettaglio come procede il processo di unificazione propriamente detto. Innanzitutto, c'è da osservare che quasi mai esso riguarda due sole persone, cioè soggetto e candidato; infatti, nella grande maggioranza dei casi vi sono altre persone collegate a esse da vincoli di parentela. Per esempio, se il sistema decide di unificare due persone di nome Antonio che hanno entrambe il padre di nome Vittorio, è ovvio che oltre ai due Antonio dovranno essere unificati anche i due Vittorio. Quindi sebbene l'unificazione inizi sempre da due persone, successivamente ne coinvolge altre, a volte molte altre.

Per comprendere come funziona il processo esamineremo qualche esempio pratico e, a questo scopo, utilizzeremo il file di documentazione che viene prodotto dal programma stesso via via che procedono le unificazioni. Cominciamo da un caso semplice: l'unificazione tra Vittorio 6081 e Vittorio 387[27] riportata nelle figg. 5 e 6.

Cominciamo dalla prima riga di fig. 6 che contiene il codice «(2)»; ciò indica che si tratta del secondo "giro" del programma[28]; «-1629-» indica il numero di persone che fino a questo punto sono state eliminate dall'archivio a causa delle precedenti unificazioni. La parola «esamino» indica che in questo momento Vittorio 6081 è stato scelto come soggetto per l'unificazione. Successivamente vediamo l'elenco dei candidati, già ordinato in base al punteggio; si nota che il primo della lista, con 4 punti, è Vittorio 387 che viene dunque scelto per l'unificazione. In fondo alla riga che lo riguarda troviamo il codice «a2k1r1» che indica in maniera sintetica come è stato calcolato il punteggio[29].

[27] I numeri 6081 e 387 indicano il codice progressivo che è assegnato automaticamente dai programmi costruttori a ogni persona per distinguerla dalle altre.

[28] Nel primo vengono prese in considerazione solo le unificazioni con intervento manuale, di cui parleremo in seguito.

[29] A2 indica 2 punti per il fatto che Vittorio è un nome non comune; k1, un punto perché soggetto e candidato hanno entrambi un figlio con lo stesso nome (Cerbone). Il punto corrispondente a r1 rappresenta un criterio un po' particolare, utilizzato per tenere conto del fatto che le persone nate prima del 1700 hanno nel database meno antenati rispetto a quelle nate dopo e quindi, potenzialmente, meno punteggio a parità di condizioni.

Fig. 5. Schema dell'unificazione di Vittorio 6081 e Vittorio 387

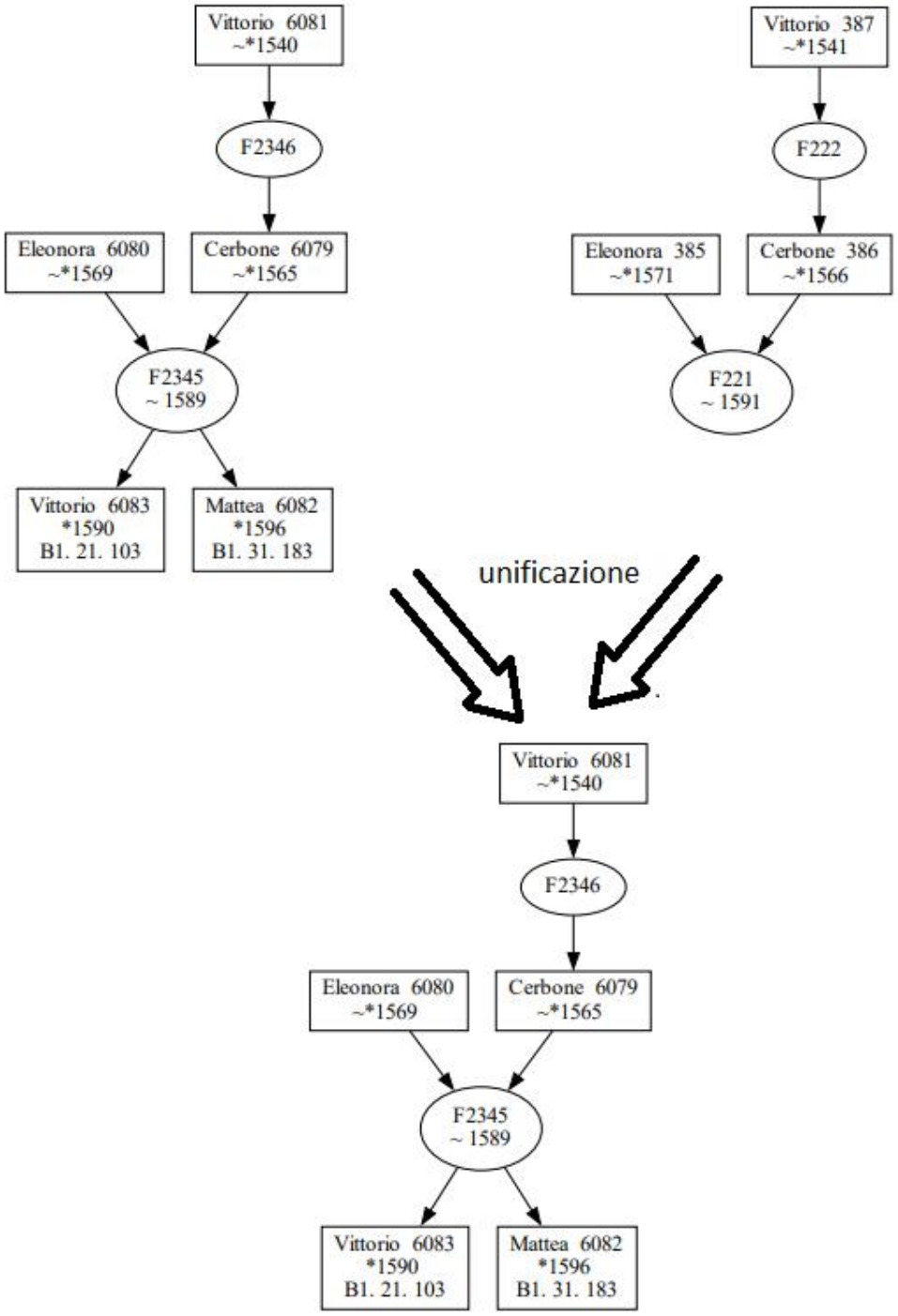


Fig. 6. Documentazione prodotta durante l'unificazione di Vittorio 6081 e Vittorio 387

```
*************

(2) - 1629 - ESAMINO: M*Vittorio*1540****6081****Cerbone***
CANDIDATI:
04^M*Vittorio*1541****387****Cerbone***^A2K1R1
-1^M*Vittorio*1587**Trolio**42289**Lorenza*****^A2E-25
-1^M*Vittorio*1585*Sardi***43261**Giacoma*****^A2B-1E-25R1
-1^M*Vittorio*1582**Domenico**25695**Lorenza**Bartolomea,Giovanni***^A2E-25
-1^M*Vittorio*1581**Trolio**26086**Lorenza*****^A2E-25
SCELTO:
 04^M*Vittorio*1541****387****Cerbone***^A2K1R1
 Controlli superati
 Non ho identificato antenati comuni di Vittorio 387 e Vittorio 6081
 >> unisco Vittorio 387 e Vittorio 6081: Vittorio 6081 resta e Vittorio 387 sparisce
    Ho trasferito il figlio Cerbone 386 nella famiglia F2346 di Vittorio 6081
    Tra i figli di Vittorio 6081 sono presenti gli omonimi Cerbone 6079,Cerbone 386
    -- Cerbone 6079 e Cerbone 386 verranno uniti
    Ho eliminato Vittorio 387
 >> unisco Cerbone 6079 e Cerbone 386: Cerbone 6079 resta e Cerbone 386 sparisce
    Tra i coniugi di Cerbone 6079 e Cerbone 386 ci sono gli omonimi Eleonora 6080 e Eleonora 38
 Non ho identificato antenati comuni di Eleonora 6080 e Eleonora 385
    Ho eliminato Cerbone 386
 >> unisco Eleonora 6080 e Eleonora 385: Eleonora 6080 resta e Eleonora 385 sparisce
    Ho eliminato Eleonora 385

*************

(2) - 1632 - ESAMINO: M*Vittorio*1581**Trolio**26086**Lorenza*****
CANDIDATI:
07^M*Vittorio*1587**Trolio**42289**Lorenza*****^Q3A2G2
```

La frase «controlli superati» ci informa che i controlli preventivi standard sono stati superati e quindi l'unificazione può procedere; la frase «non ho identificato antenati comuni di Vittorio 387 e Vittorio 608» indica che non è possibile risalire all'indietro nell'albero genealogico e che l'unificazione può iniziare con Vittorio 387 e Vittorio 6081[30].

Il sistema decide poi quale delle due persone mantenere e quale eliminare; la scelta viene fatta in modo da minimizzare il lavoro necessario. Si passa quindi all'unione vera e propria, che in pratica consiste nel trasferire in maniera coerente tutti gli attributi dalla persona da eliminare a quella che resterà; successivamente quella da eliminare verrà cancellata dal database. Vengono così trasferite tutte le citazioni con i riferimenti ai documenti e le eventuali date conosciute con esattezza (nascita, matrimoni e morte); nello stesso modo vengono trasferite la o le famiglie discendenti. Più precisamente: se la persona che resta non ha già una famiglia, riceve *in toto* la famiglia della persona da eliminare; se invece ha già una famiglia, in un primo tempo verranno aggiunti a questa famiglia i figli e il coniuge dell'altra e, successi-

[30] Il processo di unificazione deve iniziare sempre dal più alto antenato comune; ciò significa che se soggetto e candidato avessero avuto entrambi un padre o un nonno, l'unificazione sarebbe iniziata da questo.

vamente, il programma analizzerà gli eventuali omonimi per stabilire se si tratta della stessa persona o di persone diverse[31].

Nell'esempio che stiamo esaminando abbiamo proprio quest'ultimo caso: a seguito dell'unione dei due Vittorio, il figlio Cerbone 386 viene aggiunto alla famiglia F2346 di Vittorio 6081. In questa famiglia si trovano ora due figli omonimi e il sistema decide che si tratta della stessa persona[32]; pertanto anche Cerbone 6079 e Cerbone 386 dovranno essere uniti. Nello stesso modo, quando i due Cerbone vengono uniti, la moglie Eleonora 385 viene aggiunta alla famiglia F2345 di Cerbone 6079, dove già è presente Eleonora 6080, e il sistema decide anche in questo caso che si tratta della stessa persona.

Come si può constatare, alla fine del processo sono state eliminate dal database in tutto tre persone e due famiglie; inoltre, anche se ciò non è visibile dallo schema di fig. 6, Vittorio 6081 adesso ha acquisito una citazione, ossia il riferimento a un documento, che originariamente apparteneva a Vittorio 387. Vediamo ora un esempio un po' più complesso che comprende, tra le altre cose, la cosiddetta "propagazione del cognome". La situazione, prima dell'unificazione, è schematizzata in fig. 7.

A un certo punto il sistema valuta che siano da unificare Virgilio 4502 e Virgilio 4528; quello che c'è di particolare è il fatto che, in questo caso, solo il primo dei due ha il cognome. La situazione è emblematica perché i cognomi tendono a comparire intorno alla metà del XVII secolo; quindi, può succedere che una stessa persona non abbia il cognome in una registrazione più antica e poi lo acquisisca in una più recente. In casi come questo il sistema provvede automaticamente a propagare[33] il cognome.

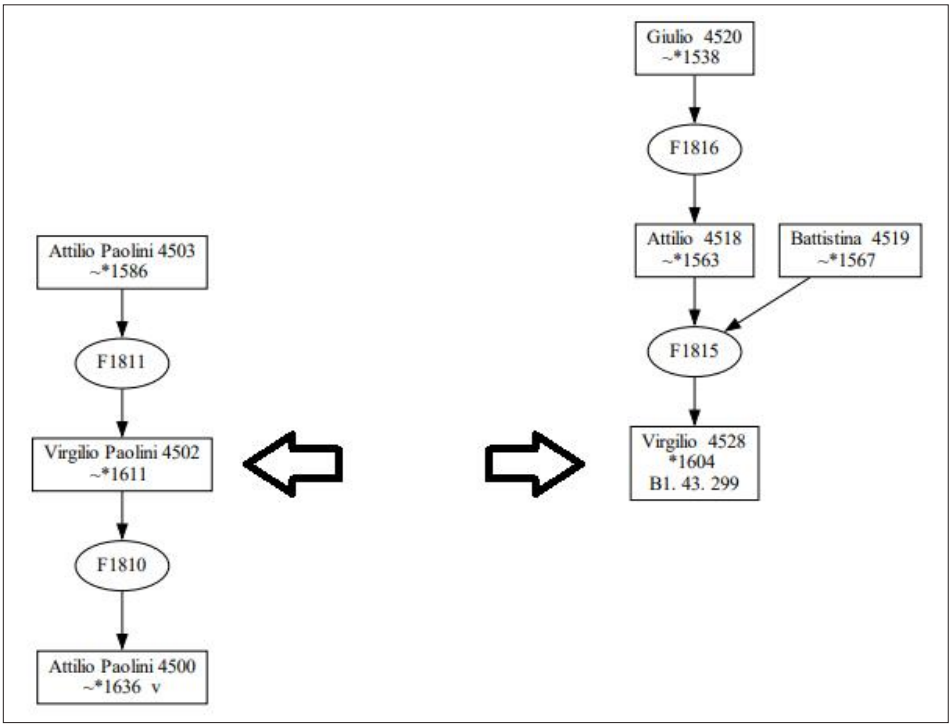
In fig. 8 vediamo come procede l'unificazione, basandoci ancora una volta sulla documentazione prodotta dal sistema. La scelta del candidato avviene nello stesso modo visto nell'esempio precedente; tuttavia, questa volta il sistema rileva la presenza di un antenato comune, Attilio, quindi l'unificazione inizierà da lui. È qui che si riscontra la necessità di propagare il cognome: l'albero genealogico viene risalito fino all'antenato più alto, cioè Giulio, e da lui, ricor-

[31] È importante sottolineare che in tutto questo processo non viene persa alcuna informazione, ma si ha piuttosto un miglioramento della precisione. Per fare un esempio: prima dell'unificazione uno dei due potrebbe avere data di nascita stimata e l'altro data certa, con registrazione battesimale. Dopo l'unificazione rimarrà una sola persona con data di nascita certa.

[32] Se risulta che entrambi gli omonimi sono diventati adulti, necessariamente deve trattarsi della stessa persona; in questo caso entrambi sono coniugati, quindi ne consegue la decisione.

[33] La propagazione del cognome avviene in questo modo: per prima cosa il sistema ricerca l'antenato maschile di ordine più alto nell'albero genealogico e gli aggiunge il cognome; successivamente aggiunge il cognome a tutti i suoi discendenti. In alcuni casi questa operazione può riguardare molte decine di persone.

Fig. 7. Situazione prima dell'unificazione

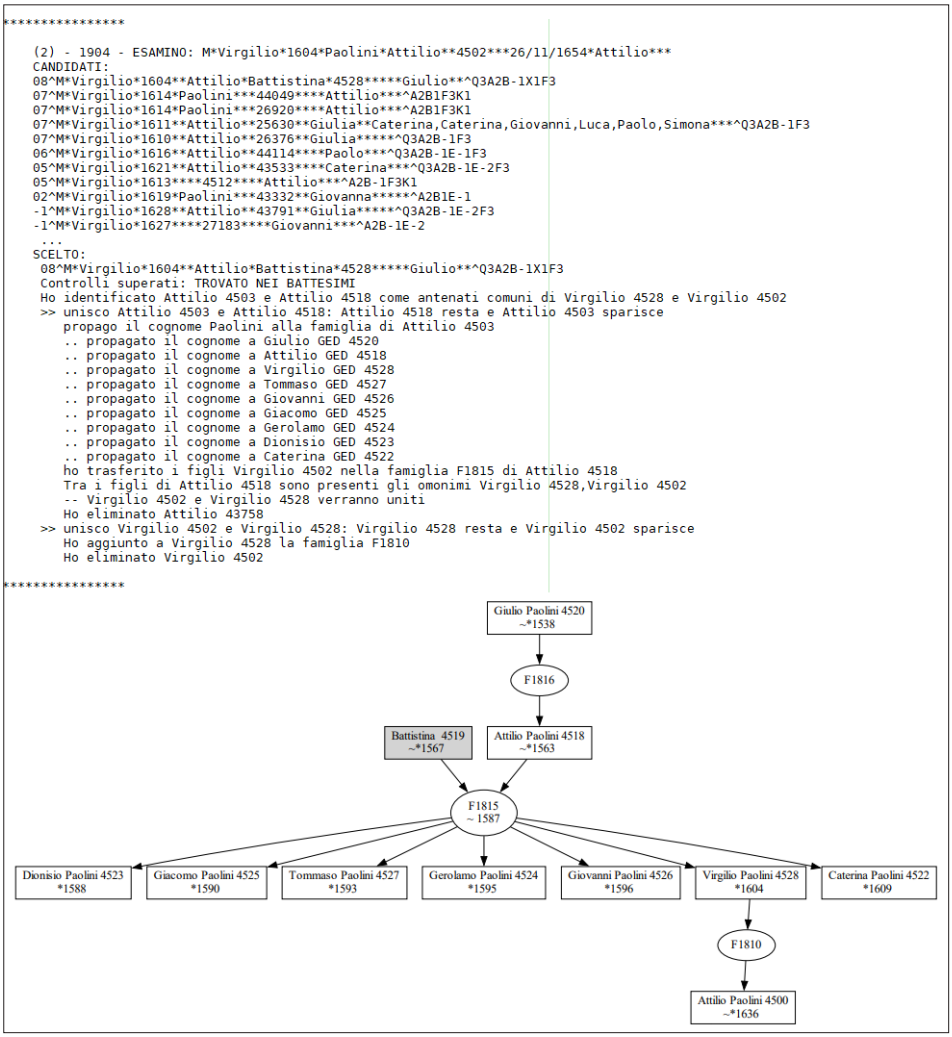


sivamente, il cognome viene propagato a tutti i suoi discendenti. Per maggior chiarezza nella stessa fig. 8 è riportato lo schema con i discendenti di Giulio.

Il lettore attento avrà osservato che il candidato scelto per l'unificazione, Virgilio 4528, ha un punteggio pari a 8, che è piuttosto elevato, ma che ve ne sono anche altri con punteggio alto; in effetti è solo una questione di "precedenza". Successivamente uno dopo l'altro anche questi verranno unificati con il soggetto Virgilio 4502, che a ogni ciclo vedrà ampliata la propria famiglia e/o il numero delle citazioni che lo riguardano. A questo proposito occorre osservare che ci sono due tipi di unificazioni: quelle dirette, in cui una persona è coinvolta come soggetto o come candidato, e quelle indirette, nelle quali la persona è coinvolta in qualità di parente o affine di un'altra che è stata unificata direttamente.

Tenuto conto di ciò, risulta che talvolta una singola persona può subire decine di unificazioni, ed è interessante notare che ognuna di esse lascia una traccia nell'elenco delle sue citazioni. A titolo di esempio in fig. 9 è riportato il frammento del database Gedcom relativo a un certo Virgilio 4528 che, come si può vedere, riporta ben 22 citazioni; questo significa che direttamente o indirettamente questa persona è stata coinvolta in 21 unificazioni.

Fig. 8. Unificazione con propagazione del cognome



In generale il programma di unificazione è in grado di gestire situazioni molto complesse, che ben possono essere esemplificate dal caso di Benedetto Sardi, rappresentato in fig. 10, che ebbe ben quattro matrimoni.

Per concludere la panoramica sul software Sage, esaminiamo brevemente gli strumenti previsti a garanzia della coerenza del database. Abbiamo già accennato a un sistema di controllo *a priori* che interviene subito prima che l'unificazione abbia luogo, quando già sono stati superati tutti i controlli standard di compatibilità. Per comprendere meglio quale situazione può dare luogo a errori di questo tipo, conviene rifarsi all'esempio di fig. 11.

Fig. 9. Elenco delle citazioni relative a Virgilio 4528

```
0 @I4528@ INDI
1 NAME Virgilio \Paolini\
1 SEX M
1 BIRT
2 DATE 21/04/1604
1 DEAT
2 DATE 26/11/1654
1 CITA
2 NOTE 1604*batt *B1. 43. 299
2 NOTE 1635*sposo*M1. 47. 178
2 NOTE 1636*padre batt*B1.183.1067
2 NOTE 1639*padre batt*B1.193.1138
2 NOTE 1641*padre batt*B1.213.  58
2 NOTE 1644*padre batt*B1.220. 112
2 NOTE 1647*padre batt*B1.228. 186
2 NOTE 1647*padre def*D1. 41_717
2 NOTE 1651*padre batt*B1.240. 285
2 NOTE 1654*defunto*D1. 47_824
2 NOTE 1656*marito defunta*D1. 48_844
2 NOTE 1661*nonno pat batt*B1.264. 516
2 NOTE 1663*nonno pat batt*B1.268. 555
2 NOTE 1664*padre sposo*M1. 61. 229
2 NOTE 1665*nonno pat batt*B1.271. 598
2 NOTE 1665*nonno pat def*D1. 54_968
2 NOTE 1667*nonno pat batt*B1.276. 651
2 NOTE 1667*padre def*D1. 56_1001
2 NOTE 1669*nonno pat batt*B1.282. 707
2 NOTE 1677*nonno pat batt*B1.311. 899
2 NOTE 1679*nonno pat batt*B1.207.1242
2 NOTE 1683*nonno pat batt*B1.322. 972
1 FAMC @F1815@
1 FAMS @F9535@
```

Fig. 10. I quattro matrimoni di Benedetto Sardi

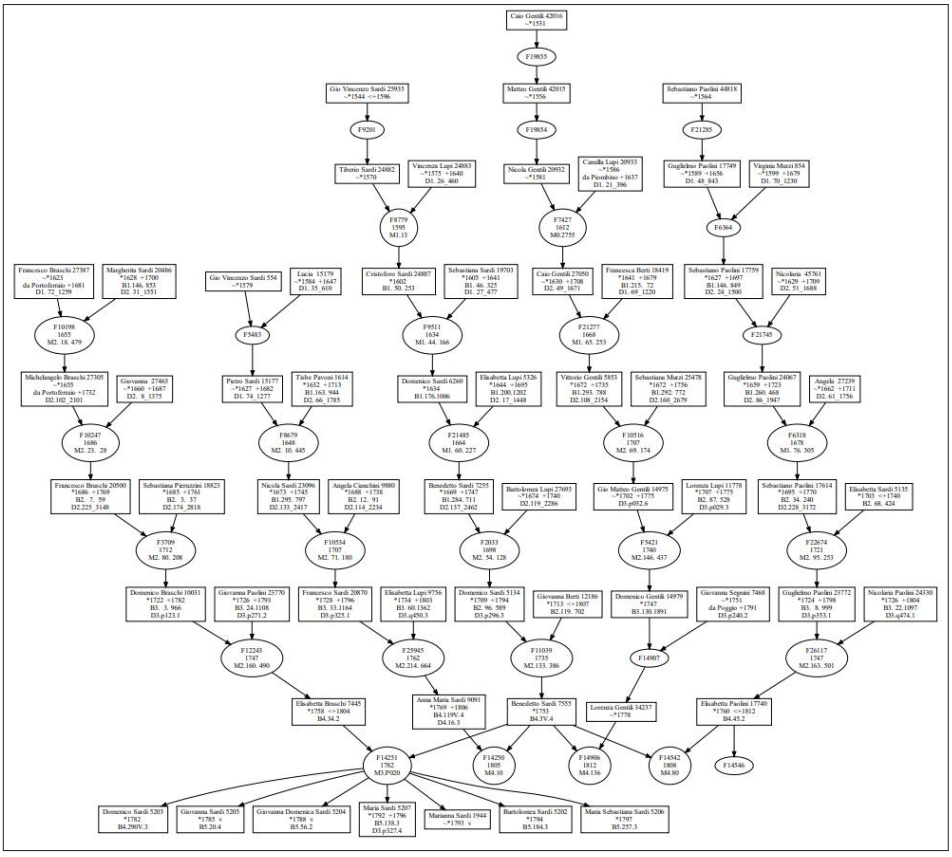
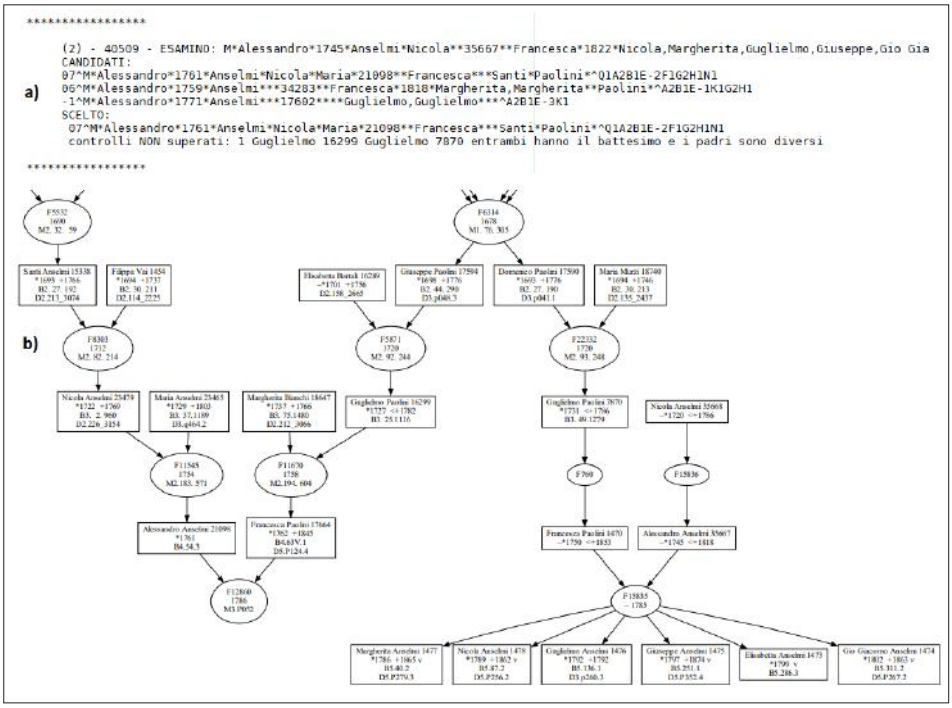


Fig. 11. Esempio di unificazione bloccata automaticamente perché errata



Nella figura 11a) è riportato il frammento di documentazione da cui si vede che il sistema stava per procedere all'unificazione di Alessandro 35667 e Alessandro 21098 con un punteggio di 7, che è piuttosto alto, ma si è bloccato perché ciò avrebbe provocato erroneamente l'unione di Guglielmo 16299 e Guglielmo 7870 che sono due persone diverse.

La situazione appare subito chiara esaminando gli schemi famigliari. Alessandro 35667 e Alessandro 21098 corrispondono certamente alla stessa persona, perché si ha coincidenza nel nome del padre, della moglie e del suocero. Tuttavia, esistono due cugini primi di nome Guglielmo e durante qualche precedente unificazione, il sistema aveva erroneamente assegnato Francesca 1470 (moglie di Alessandro 35667) come figlia di Guglielmo 7870. Tale scelta, benché errata, al momento in cui è stata fatta era legittima in quanto risultava perfettamente compatibile con le informazioni disponibili. In realtà il "giusto" Guglielmo era il cugino 16299, ma il sistema ne ignorava l'esistenza.

Situazioni di questo tipo non sono rare dato che l'abitudine diffusa di assegnare ai figli i nomi dei nonni, portava ad avere spesso dei cugini omonimi. Il sistema è in grado di risolvere automaticamente l'ambiguità se sono citate le madri ma, quando come in questo caso sono omesse, è indispensabile l'in-

Fig. 12. Esempio di rollback

```
*****************

     (1) - 518 - ESAMINO: M*Francesco*1741*Vai*Giovanni**9814**Maria**Alessandro,Alessandro,Giovanni,Luigi,Pi
     CANDIDATI:
     25^M*Francesco*1743*Vai*Giovanni**29469**Maria***Francesco*Braschi*^MAN-xmat
     SCELTO:
      25^M*Francesco*1743*Vai*Giovanni**29469**Maria***Francesco*Braschi*^MAN-xmat
      Controlli superati: TROVATO NEI MATRIMONI
      Ho identificato Giovanni 29470 e Giovanni 9816 come antenati comuni di Francesco 29469 e Francesco 9814
      >> unisco Giovanni 29470 e Giovanni 9816: Giovanni 29470 resta e Giovanni 9816 sparisce
         ho anticipato l'anno di nascita stimata di Giovanni 29470 dal 1718 al 1716
         Ho trasferito il figlio Francesco 9814 nella famiglia F11620 di Giovanni 29470
         Tra i figli di Giovanni 29470 sono presenti gli omonimi Francesco 29469,Francesco 9814
         -- Francesco 9814 e Francesco 29469 verranno uniti
         Ho eliminato Giovanni 9816
      >> unisco Francesco 9814 e Francesco 29469: Francesco 29469 resta e Francesco 9814 sparisce
         Tra i coniugi di Francesco 29469 e Francesco 9814 ci sono gli omonimi Maria 29473 e Maria 9815 che v
         Ho aggiunto a Francesco 29469 i figli Alessandro 9819,Alessandro 9820,Giovanni 9821,Luigi 9822,Pietr
         Tra i figli di Francesco 29469 sono presenti gli omonimi Alessandro 9819,Alessandro 9820
         -- Alessandro 9819 e Alessandro 9820 sono persone diverse
      Ho identificato Francesco 29475 e Francesco 9818 come antenati comuni di Maria 29473 e Maria 9815
         Ho eliminato Francesco 9814
GRAVE ERRORE: Francesco 29469 risulta sposato a 13 anni
INIZIO ROLLBACK
    > Francesco Vai 9814 ripristinato
      Maria Braschi 9815: modifica annullata
    > Giovanni Vai 9816 ripristinato
      famiglia F3640 ripristinata
      Alessandro Vai 9819: modifica annullata
      Alessandro Vai 9820: modifica annullata
      Giovanni Vai 9821: modifica annullata
      Luigi Vai 9822: modifica annullata
      Pietro Vai 9823: modifica annullata
      Sebastiano Vai 9824: modifica annullata
      famiglia F3639 ripristinata
      Giovanni Vai 29470: modifica annullata
      Francesco Vai 29469: modifica annullata
      famiglia F11623: modifica annullata
      files sequenziali aggiornati
ROLLBACK CONCLUSO

*****************
```

tervento manuale. Vediamo ora gli altri due sistemi predisposti per la verifica della coerenza del database. Il primo interviene subito dopo che è avvenuta una singola unificazione e cerca di evidenziare, a posteriori, eventuali problemi sfuggiti ai controlli preventivi. Gli errori di questo tipo sono piuttosto rari, circa 5 ogni 1000 unificazioni, ma è necessario correggerli immediatamente per evitare che le incoerenze si propaghino a macchia d'olio compromettendo in maniera imprevedibile l'affidabilità complessiva del database. Quando riconosce la presenza di un errore, il sistema risponde effettuando un *rollback*, cioè ripercorrendo all'indietro e annullando tutte le modifiche apportate al database a seguito dell'unificazione errata, in modo da ripristinare la situazione precedente. Ciò è possibile perché il sistema tiene traccia, in un apposito file, di ogni modifica effettuata al database.

Nella figura 12 è riportata la documentazione prodotta in uno di questi casi. La procedura è stata attivata dopo che il sistema ha riscontrato che a seguito dell'unificazione Francesco 29469 risultava sposato a 13 anni, età che è considerata inaccettabile per una persona di sesso maschile e quindi indicativa di una condizione di errore. A quel punto è possibile che il sistema riesca a trovare autonomamente una soluzione diversa, oppure che sia necessario un intervento manuale.

L'ultimo strumento a cui si accennerà è un autonomo programma di verifica complessiva di coerenza del database che viene lanciato dopo la conclusione di tutte le unificazioni; i controlli sono molto approfonditi e servono a evidenziare gli errori sfuggiti agli altri strumenti. Anche queste anomalie sono rare, circa 7 ogni 1000 unificazioni e la loro correzione richiede necessariamente un intervento manuale.

Ma quali sono le cause di questi errori? Per quanto si può ricavare dall'esperienza, le fonti di errore sono diverse; tra le più comuni possiamo citare gli errori nella decodifica dei documenti, i cognomi riportati con diverse grafie (soprattutto per i forestieri), i luoghi di provenienza dei forestieri che in registrazioni diverse sono riportati in maniera diversa (per es. in una registrazione si trova Camogli e in un'altra Genovese), i nomi plurimi che vengono usati in maniera discordante. Alcuni errori hanno origine nella fonte documentale stessa e vengono evidenziati solo grazie al programma di verifica del database: tipico è il caso di errore nel nome di un genitore o di un nonno morto molto prima della data dell'evento oggetto della registrazione. Infine, qualche errore è risultato essere causato anche da bug nel software che, con le sue circa 10.000 righe di programma Python, nel corso degli anni è diventato un sistema piuttosto complesso.

5. L'intervento manuale. Come ricordato in precedenza la strategia utilizzata con il software Sage, di tipo iterativo, nel senso che dopo ogni *run* occorre intervenire manualmente per correggere gli eventuali problemi e/o per migliorare i risultati.

Il programma di verifica del database fornisce indicazioni piuttosto precise su dove occorre intervenire per correggere gli errori; inoltre elenca tutta una serie di anomalie che potrebbero essere errori oppure no. Per fare qualche esempio, segnala le persone che risultano avere due coniugi con lo stesso cognome ma nomi diversi; occorre quindi verificare se si tratta dello stesso coniuge che compare con nomi diversi (es. Maria e Maria Giovanna o Giovanna Maria), oppure di una sorella o un fratello sposato dopo la morte del primo coniuge, oppure se è una pura coincidenza. Analogamente segnala le persone che hanno due coniugi con lo stesso nome ma cognomi diversi; anche in questo caso potrebbe trattarsi di persone realmente diverse oppure della stessa persona che viene citata con cognome trascritto in maniera diversa.

Gli interventi manuali, come già detto, vengono fatti direttamente sul foglio elettronico e sono sostanzialmente di tre tipi: la correzione (per esempio un cognome errato), l'integrazione (per esempio l'aggiunta di un genitore non presente in una registrazione di matrimonio, in modo da "indirizzare" opportunamente il sistema) e la forzatura. Quest'ultima è la possibilità di forzare

manualmente l'unione di due persone, segnalando al programma di unificazione che la persona che compare in due o più registrazioni diverse è in realtà la stessa persona. Ciò si ottiene molto semplicemente sfruttando la coincidenza della data di nascita; per esempio, se si vuole fare in modo che una persona che compare come sposo in una registrazione di matrimonio venga unificata con quella che compare in un battesimo, è sufficiente indicare la sua data di battesimo in corrispondenza di un apposito campo ausiliario previsto nella riga che contiene la registrazione del matrimonio. In tutti i casi in cui si interviene manualmente sui dati del foglio elettronico, è buona norma lasciare traccia delle modifiche nell'apposito campo previsto per le note, in modo da poter da un lato ricostruire il dato realmente presente sul documento e dall'altro, se necessario, poter correggere o eliminare la modifica stessa.

Prima di concludere cercheremo di dare una valutazione della "bontà", ossia dell'affidabilità complessiva del database, tenendo conto che, come detto in precedenza, il processo di affinamento è un processo in divenire. Non sono a conoscenza di strumenti di misura oggettivi e condivisi da utilizzare a questo scopo, per cui posso solo citare dei parametri empirici. Per esempio, attualmente nel database sono presenti 19216 persone; considerando solo quelle nate dopo il 1586[34], e tolti i forestieri, ne rimangono 16867. Di queste è stato possibile identificare la registrazione battesimale per 12532 persone, cioè il 74%; per 5569, pari al 33%, sono stati identificati battesimo e sepoltura e per 2073, pari al 12%, battesimo, matrimonio e sepoltura. Un altro parametro empirico, a mio parere abbastanza significativo, che può dare un'idea del livello complessivo di connessione delle reti famigliari, è la percentuale di persone senza famiglia di origine[35]; durante gli anni trascorsi nel processo di affinamento del database, esso è passato da un iniziale 33% all'attuale 12%. Una parte di queste persone senza padre compaiono in registrazioni molto antiche e quindi rappresentano in qualche modo i punti di partenza della rete genealogica; le restanti costituiscono invece delle discontinuità, dei veri e propri "buchi nella rete", che occorre pazientemente rammendare uno a uno come fanno i pescatori.

6. *Prospettive future*. Al momento è in corso una profonda ristrutturazione di tutto il software che dovrebbe renderne più semplice l'utilizzo da parte di altri ricercatori e, allo stesso tempo, ridurre di uno o due ordini di grandezza il tempo di calcolo necessario. I ricercatori eventualmente interessati a utilizzare il software possono fin da ora contattare l'autore che intende renderlo disponibile in forma *open source* Gnu per scopi di ricerca.

[34] Anno in cui iniziano le registrazioni dei battesimi.
[35] Esclusi i forestieri e i figli di genitori ignoti.