Tesis de Licenciatura

Una herramienta computacional para la reconstrucción de genealogías históricas

Carlos Diuk
cd4k@dc.uba.ar

Directores: Dr. Pablo Jacovkis y Dr. Enrique Tandeter

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Agosto de 2003

Figura 1: Dibujo de Pérez Bocanegra, incluído en tratado pastoral de 1631, con el objeto de ilustrar las prohibiciones matrimoniales basadas en el grado de consanguinidad.

Capítulo 1

Introducción

El presente trabajo presenta una metodología general, acompañada por un conjunto de herramientas informáticas, para la reconstrucción de genealogías históricas basada en registros parroquiales (bautismos, matrimonios, defunciones) o estatales (censos).

El problema principal consiste en identificar, en diferentes fuentes históricas, referencias a un mismo individuo real y vincular, por lo tanto, dichas fuentes.

Las herramientas desarrolladas y la metodología propuesta fueron aplicadas exitosamente en la reconstrucción de genealogías parciales de los habitantes de los pueblos de Sacaca y Acasio (Alto Perú) durante el período de 120 años que se extiende entre 1690 y 1810. Como fuente, se utilizaron las 11750 actas matrimoniales, debidamente digitalizadas, de estos dos pueblos.

Para dar cuenta del punto donde se inserta el presente trabajo comenzaremos por un breve recorrido historiográfico. Nos interesa en particular recorrer los cambios que se produjeron en el campo de la Historia a partir del siglo XIX y que, con la aparición y el desarrollo de los métodos cuantitativos en el siglo XX, convirtieron a la computadora en una herramienta fundamental para un nuevo grupo de historiadores.

1.1 Profesionalización, debate y crisis. La Historia y las Ciencias Sociales

El temprano siglo XIX observó un cambio fundamental en la forma en que la Historia se investigaba, leía y enseñaba. Básicamente, podemos caracterizarlo como un período de profesionalización, donde la Historia entra definitivamente al mundo universitario y busca su espacio como Ciencia por derecho propio, seguramente empujada por las corrientes cientificistas que acompañaron el auge de las ciencias naturales hasta entonces.

Dentro de este proceso, podemos identificar como punta de lanza a la escuela alemana, y principalmente a Leopold von Ranke desde la Universidad de Berlín. A partir de 1848 en Alemania, y alrededor de 1870 en la mayor parte de Europa, Estados Unidos y Japón, la historia se profesionaliza, se institucionaliza, define su método y vive un verdadero auge con la aparición de revistas especializadas, dedicadas a difundir las nuevas metodologías de esta nueva ciencia escolástica (ver [15]).

A pesar de los cambios en el método, hasta bien avanzada la segunda mitad del siglo XIX podemos caracterizar a la Historia como un campo preocupado esencialmente por narrar los acontecimientos políticos, bélicos, institucionales y diplomáticos de las naciones. Su centro era la Nación y sus "héroes", junto con los avatares políticos y diplomáticos de los que formaran parte. Aunque existieran antecedentes de otro tipo de historia en el siglo XVIII, una historia que sus mismos autores llamaban "historia de la sociedad", la profesionalización encarada por Ranke y sus discípulos la marginó completamente, no considerándola una disciplina académica y prácticamente estigmatizando a sus cultores como *diletantes* (ver [5]).

Esta concepción comienza a ser cuestionada, y a fines del siglo XIX surgen las primeras voces que proponen un acercamiento de la historia hacia las ciencias sociales, otorgándole un rol mayor en su objeto de estudio a la sociedad, la economía y la cultura.

En Alemania, podemos citar como una de estas primeras voces a Karl Lamprecht y su primer tomo de la *Historia de Alemania* [19], aparecido en 1891. Lamprecht cuestiona el rol central que la Academia le asignaba hasta el momento al Estado, y su concentración en los eventos y los grandes protagonistas. Tomando como ejemplo a las ciencias naturales, que ya hacía tiempo habían dejado atrás su caracter meramente descriptivo de fenómenos aislados para volcarse a las explicaciones de caracter general, Lamprecht presenta un enfoque amplio, otorgándole gran importancia al contexto social, económico y político de la época en estudio. Lamprecht oponía la historia política, que era una historia de individuos, a la historia cultural o económica, que era la historia del pueblo. A pesar de su buena aceptación en el público general, Lampretch enfrentó fuertes resistencias y críticas desde la Academia alemana.

Mientras tanto, en Francia, principalmente a partir del fin de la guerra franco-prusiana, algunos historiadores intentaban recuperar terreno en su atrasada carrera historiográfica, imitando el modelo de historia de la escuela alemana. Entre ellos mencionaremos a Lavisse y Seignobos, contra quienes luego se alzarían críticas similares a las de Lamprecht.

Es en esta época donde en Francia podemos observar un gran auge de la sociología, principalmente de la mano de Emile Durkheim. En 1888, Durkheim publica su "Curso de Ciencia Social" [6], donde cuestiona el caracter científico de la historia por carecer justamente de una preocupación por obtener conceptos generales capaces de validación empírica, y por estar concentrada sobre lo particular. François Simiand, economista y discípulo de Durkheim, diagnostica que la Historia, adoradora de la cronología, el individuo y la política, está incapacitada para convertirse en una verdadera ciencia social. Al mismo tiempo, plantea que la historia económica, ella sí preocupada por los modelos y las cantidades, es una subdivisión de la Historia que puede adquirir el ansiado —a su criterio—, status de ciencia social.

Y es aquí, entre 1900 y 1930 donde, desde los intentos de Lavisse y Seignobos, pasando por la crítica de Durkheim y Simiand, la escuela francesa, hasta entonces notablemente atrasada con respecto a su par alemana, comienza a emparejar la carrera hasta liderar definitivamente el campo epistemológico con la fundación de *Annales* en 1929.

Podemos mencionar, por último, que un proceso similar ocurría en Estados Unidos con el surgimiento de los *New Historians*, también conocidos como *Progressive Historians*.

Veremos ahora cómo estos proceso y estas críticas desembocan, volviendo a Francia, en la aparición del movimiento de *Annales*.

1.2 Los *Annales*

El *movimiento de Annales* —sus miembros renegaban del mote de *escuela*— refiere al conjunto de ideas y personas que giró en torno a una revista, *"Annales d'histoire économique et sociale"*, fundada y dirigida en sus primeros años por Lucien Febvre y Marc Bloch. La revista ejerció una influencia fundamental sobre el campo de la Historia a lo largo de su existencia, desde 1929 hasta la actualidad, tanto en Francia —probablemente relanzando y llevando a la vanguardia a la escuela francesa— como en el resto del mundo.

Revisaremos brevemente la historia de los *Annales* sin un espíritu exhaustivo ni extremadamente cronológico, sino con el objetivo de comprender cómo se llega al auge de la cuantificación en historia, al diálogo con otras disciplinas, y en particular a la incorporación de los métodos informáticos en el campo.

1.2.1 Los comienzos

Annales nace, quizás, de la oportunidad.

Con epicentro en la Sorbona, bajo los dominios de Ernest Lavisse y su *"L'histoire de France"*, la historia cobra gran auge en Francia en los primeros años del siglo XX y, a pesar de cierta apertura hacia consideraciones culturales y geográficas, sigue siendo una historia eminentemente política, constitutiva de *lo nacional* y sus valores (ver [4, 5, 25]).

Es entonces en la periferia, y en la oportunidad, donde se constituye el nuevo movimiento. Luego de la Primera Guerra Mundial, Francia recupera la ciudad de Estrasburgo y se reorganiza —aunque refunda— su Universidad. Este medio favorecía las innovaciones intelectuales, el intercambio de ideas y la ruptura de las rígidas fronteras disciplinarias impuestas en ámbitos ya establecidos. En palabras de André Burguière (ver [4]), la Universidad de Estrasburgo ofrecía a Febvre y Bloch "un vivero intelectual prácticamente sin igual en Francia: un medio habituado a los debates interdisciplinarios (como aquellas reuniones de los sábados en las cuales geógrafos, sociólogos, lingüistas e historiadores confrontaban sus aproximaciones) y sensibilizado a los temas que iban a definir la identidad científica de la revista". Y es en la nueva Universidad de Estrasburgo donde Febvre y Bloch comparten cargos, donde se conocen y entablan una importante amistad a partir de 1920.

Lucien Febvre se forma en la Ecole Normale Supérieure, donde ingresa en 1897. Es allí donde se ve fuertemente influenciado por cuatro de sus profesores: el geógrafo Paul Vidal de la Blache, sumamente interesado en colaborar con historiadores y sociólogos; el filósofo y antopólogo Lucien Lévy-Bruhl, dedicado al estudio de la "mentalidad primitiva"; el historiador del arte Emile Mâle; y el lingüista Antoine Meillet, discípulo de Durkheim e interesado en los aspectos sociales del lenguaje (ver [5]). Esta influencia múltiple, interdisciplinaria y con un enfoque cultural se completa con la influencia del socialismo de Jean Jaurès, que puede apreciarse en la tesis doctoral de Febvre. En su estudio sobre su región de origen —el Franco Condado—, bajo el dominio de Felipe II en el siglo XVI, Febvre ubica su mirada sobre la lucha entre dos clases: la nobleza en

decadencia y la burguesía ascendente. Aunque pueda verse este enfoque como un análisis cercano al marxismo, Febvre se diferencia al señalar que su mirada no concibe la lucha entre los grupos como "mero conflicto económico sino también como conflicto de ideas y sentimientos" (ver [7]). Por último, Febvre introduce en sus trabajos una mirada geográfica, donde toda descripción de una situación histórica comienza introduciendo una descripción geográfica de la región donde ocurre.

Marc Bloch también se forma en la Ecole Normale Supérieure, y también recibe influencias de Meillet y Lévy-Bruhl, aunque su mayor influencia proviene de Emile Durkheim, en ese momento profesor en la Ecole. Bloch se especializa en la Edad Media y también le otorga gran importancia a la geografía histórica, aunque su mayor interés es por la sociología y, más adelante, por la historia económica.

Ambos, Febvre y Bloch, pensaban claramente de una manera interdisciplinaria.

1.2.2 El "programa" de Febvre y Bloch

Annales poseía un estilo polémico y directo, que contrastaba con la habitual prudencia y mesura universitarias, que tantas veces conspiran con el debate de ideas y la puesta en juego de posturas epistemológicas. Este estilo generaba un doble proceso: el de la obtención de enemigos, al mismo tiempo que el de la consolidación de un espíritu de grupo.

Annales establece la necesidad de estudiar la historia de los grupos sociales y las fuerzas colectivas. Por otro lado, como ya hemos mencionado, encuentra en la interdisciplina su modo de acción, nutriéndose principalmente de tres corrientes intelectuales: la escuela geográfica de Vidal de la Blache —quien fuera maestro de Febvre y Bloch en la Ecole—, la sociología durkheimiana y el movimiento creado por Henri Berr en torno a la revista *Revue de Synthèse*. Es Henri Berr quien introduce al campo de la historia una visión psicológica, buscando llevar la historia de las ideas hacia una historia de las representaciones mentales y los fenómenos de psicología colectiva (ver [4]). En relación a este último punto, y como un nuevo contraste con la escuela marxista, nos gustaría citar un comentario de Marc Bloch: "... M. Thompson, cuyo materialismo histórico no es siempre intemperante, se esfuerza con gusto en descubrir en los movimientos religiosos de la Edad Media motivos de naturaleza económica. Yo estoy, personalmente, mucho más sorprendido por los resultados económicos de los fenómenos religiosos." (ver [3])

Mencionaremos por último algunas de las formas de aproximarse al pasado que propone *Annales*. En primer lugar, el *método regresivo*, que consiste en partir de una situación presente, por ejemplo una situación geográfica, un hábito social, etc. y remontarse en el tiempo en busca de su génesis. En segundo lugar, y esto es clave, lo que dieron en llamar la *histoire problème*, la historia de los problemas en contraposición con la historia de los acontecimientos, habitual en la vieja escuela de la historia política.

Pero su rechazo a la historia política iba más allá de una cuestión metodológica. Según los primeros *Annales*, la política y la ideología sumergen al historiador en el anacronismo, al hacerlo olvidar que está observando el pasado con la óptica de su tiempo. Critican, a su vez, el rol del historiador como custodio de la Nación y de lo nacional. En palabras de Burguière (ver [4]):

"Desatando la trama de las decisiones e intenciones de los actores que ocupan la escena política tenemos la impresión de explicar todo el movimiento de la historia pero, en realidad, no hacemos sino consolidar el discurso mitológico que sostiene nuestras representaciones políticas. Los historiadores deben entonces dejar de proporcionar argumentos a la nación (o a los gobernantes), de alimentar sus necesidades de legitimidad retrospectiva, y ocuparse de proporcionar los medios para comprender mejor, y en consecuencia también dominar mejor, los mecanismos de la realidad social".

En resúmen, en los primeros Annales ya se observa la primacía de la historia como problema en contraposición a la historia cronológica y de los acontecimientos, la búsqueda de modelos en contra de la descripción de fenómenos aislados, y la convergencia con el resto de las ciencias sociales en una invitación permanente al trabajo colectivo.

1.2.3 Fernand Braudel. La segunda generación de Annales y la historia cuantitativa

Suelen identificarse tres generaciones en el movimiento de Annales. La primera, que hemos descripto en la sección anterior, es la de su fundación, con Febvre y Bloch a la cabeza. La segunda generación comienza con la primacía de un nuevo integrante del grupo, Fernand Braudel, y sus discípulos, quienes impulsan enormemente la historia cuantitativa, punto donde se inserta nuestro trabajo. De la tercera generación no hablaremos mucho en esta introducción.

Fernand Braudel estudió en la Sorbona, y planeaba una tesis sobre Felipe II y su política exterior sobre el Mediterráneo. Sin terminarla aún, viajó a San Pablo a enseñar por un período de 2 años, y la casualidad quiso que, en su largo viaje de regreso en barco, tuviera como compañero de viaje a Lucien Febvre. Durante el viaje entablaron amistad, y al desembarcar en Europa Braudel ya había decidido que su tesis, "Felipe II y el Mediterráneo", debía llamarse "El Mediterráneo y Felipe II".

La obra se torna monumental, de unas 600 mil palabras. Está dividida en tres partes: la primera introduce el medio ambiente en el cual se desarrollará la historia, con gran nivel de detalle geográfico y de manera casi atemporal; la segunda parte se sumerge en las estructuras económicas, sociales y políticas que contextualizan los hechos; y es recién la tercera parte la que trata de los acontecimientos. Esta tercera parte es, de hecho, la que había imaginado Braudel y presentado como primera tesis, antes de conocer a Febvre y las ideas de *Annales*.

Braudel se convertirá luego, sobre todo tras la muerte de Febvre en 1956, en la figura central tanto de los Annales como de la escuela francesa en general. Tras su trabajo sobre el Mediterráneo, torna al estudio de la historia económica, y en particular del capitalismo. Es aquí donde se introduce el concepto de la *larga duración*, el estudio de las estructuras y los flujos contra el estudio de los acontecimientos, el estudio de lo global contra lo fraccionado. Comienza en este marco, el de la historia económica y la larga duración, una gran preocupación por lo cuantitativo.

Podemos citar aquí a un precursor en esta tendencia, Ernest Labrousse, que en 1933 publica su trabajo sobre la historia de los precios y los ingresos en Francia, recorriendo todo el siglo XVIII (ver [18]). Braudel, a su vez, llega a analizar curvas de precios que se extienden por más de dos siglos.

Y serán principalmente los discípulos de Braudel, durante los años '60 y '70, quienes se obsesionan verdaderamente por la cuantificación. Entre ellos, mencionaremos principalmente a Emmanuel Le Roy Ladurie.

Le Roy Ladurie, obsesionado por la cuantificación y fascinado por las posibilidades que abre la informática, en un giro seguramente exagerado y quizás demasiado optimista, llega a exclamar:

"... el historiador de mañana será programador o no será historiador." (ver [20])

1.3 La demografía histórica.

1.3.1 Los inicios. De Louis Henry a la informática.

Comenzando casi naturalmente por la historia económica, el fenómeno de la cuantifiación se extiende pronto al terreno de la historia de las poblaciones, de la demografía y de lo que podemos llamar historia social del parentesco, campo en el cual se inserta nuestro trabajo.

La demografía histórica nace en la década de 1950, como un trabajo conjunto entre demógrafos e historiadores. Entre sus pioneros es central la figura de Louis Henry, quien en esa época trabajaba en el Instituto Nacional de Estudios Demográficos de Francia, y la de Pierre Goubert.

Louis Henry venía de trabajar en los años '40 sobre poblaciones del presente, y es cuando comienza a dirigir la mirada hacia las poblaciones del pasado cuando desarrolla el método de la *reconstitución de familias*. Su interés principal refiere a temas de fertilidad en la población francesa en los siglos XVII y XVIII, para los cuales existían registros parroquiales bien conservados y bien registrados. Vinculando registros de nacimientos, matrimonios y defunciones de una región, Henry intenta construir *fichas familiares*, donde se consigna la historia demográfica de los integrantes de un núcleo familiar. En 1976 publica su *Manual de demografía histórica* (ver [12]), donde da cuenta de su método con gran nivel de detalle.

A partir de 1958, varios trabajos similares se desarrollaron en Francia y sirvieron de modelo para numerosos historiadores en el resto de Europa, Japón y América del Norte. Entre ellos, un trabajo del ya mencionado Pierre Goubert, *Beauvais et le Beauvaisis de 1600 à 1730, contribution à l'histoire sociale de la France du XVIIe* (ver [10]).

Aquí aparece un primer problema relacionado con las fuentes tradicionales de la historia, usualmente preservadas en los Archivos Nacionales. Dichos archivos, que en el caso de Europa fueron generalmente constituidos en el siglo XIX, siguen los procedimientos y criterios que reflejan la preocupación ideológica y metodológica de los historiadores de la época: la preservación de los valores nacionales, y en consecuencia la prioridad dada a las fuentes político-administrativas. Por otro lado, el archivo está pensado para testimoniar sobre los eventos, y no sobre los procesos y la larga duración. (ver [8, 18]).

La demografía histórica se vuelca entonces sobre otro tipo de fuentes: aquellas que hacen referencia a las vidas de los individuos de una época y un lugar, al "hombre común" y sus pequeños eventos. Aunque en los Archivos podían encontrarse ciertos documentos de este tipo, como los censos poblacionales, los historiadores comienzan a sumergirse en los documentos de otras fuentes, las

burocracias comunales y eclesiásticas: registros de propiedad, registros impositivos y registros parroquiales (actas de bautismo, matrimonio, defunción) son rescatados y utilizados como fuente de indagación histórica.

Es aquí, entonces, donde se hace evidente el gran aporte de la informática a partir de las segunda mitad del siglo: la posibilidad de digitalizar, ordenar, revisar y sistematizar estos documentos. Este aporte se potenció a partir de fines de los años '60, con mayores capacidades de almacenamiento y procesamiento, y sobre todo con la aparición de sistemas de bases de datos comerciales. La aparición de estas nuevas herramientas acompañaron y fomentaron el desarrollo de la cuantificación en la historia, aportando nuevas posibilidades de indagación y exigiendo el desarrollo y refinamiento de ciertas metodologías de investigación.

1.3.2 La reconstrucción de familias

Los desarrollos en la demografía, sobre todo a partir de la Segunda Guerra Mundial, comenzaron a indicar que la comprensión de los fenómenos demográficos requiere de la observación continua y sostenida en el tiempo. La vida de un individuo no puede sólo leerse a través de los "eventos demográficos" [1] en los que interviene directamente, sino también a través de aquellos en los que intervienen sus parientes y su descendencia.

Para los historiadores se torna necesario, entonces, reconstruir estos vínculos familiares y seguir su desarrollo a través del tiempo. Las fuentes a disposición se refieren, sin embargo, a momentos puntuales en el tiempo (sean censos, actas de nacimiento o registros de propiedad) y a individuos o núcleos familiares pequeños (por ejemplo, hogares). De esta necesidad surgen las diversas técnicas de *reconstitución de familias*, consistentes en vincular datos de diversas fuentes a partir de referencias nominales a individuos (lo que llamaremos, en inglés, *nominal record linkage*).

En palabras de E.A.Wrigley (ver [41]), y en términos bastante generales,

> "[nominal record linkage] is the process by which items of information about a particular named individual are associated with each other into a coherent whole in accordance with certain rules.[...]"

En la mayoría de los casos, estos trabajos son realizados por un investigador solo o que, conformando un pequeño grupo de voluntarios, examina miles de registros parroquiales, a lo largo de un siglo aproximadamente, en algún pueblo de no mucho más de 1000 habitantes. Comenzando con un acta matrimonial, el investigador rastrea el resto de los eventos relativos a cada contrayente (sus nacimientos y muertes) y el de sus hijos, y los registra en una ficha familiar única. Dichas fichas permiten luego la elaboración de estadísticas demográficas, y han servido para incrementar notablemente la comprensión de los historiadores de ciertos fenómenos de comportamiento demográfico tales como fertilidad.

Sin embargo, pronto fueron claros los límites de esta metodología. Se trata de un trabajo extremadamente grande, que requiere mucho tiempo, y que produce como resultado números relativamente bajos de familias efectivamente reconstituídas. Más aún, al trabajar sólo sobre una pequeña población, el número de personas alcanzadas se ve notablemente reducido por el efecto de las migraciones. Por lo tanto, las conclusiones que se obtienen están fuertemente

[1] Básicamente, su nacimiento, matrimonio, muerte, mudanza, cambio de actividad, etc.

sesgadas por el hecho de que sólo abarcan a personas y familias sedentarias, afianzadas en un mismo pueblo a lo largo del tiempo.

Para sobrepasar este obstáculo, se tornó necesario abarcar poblaciones cada vez mayores: centros urbanos o regiones completas. Y es aquí donde los límites de la reconstrucción manual fueron alcanzados, y se volvió necesario recurrir a la informática. Nacen entonces las técnicas de reconstrucción automática de familias.

El *Cambridge Group* es pionero en este area y en 1971 E.A. Wrigley organiza la primer conferencia, en Princeton, sobre el uso de computadoras para la vinculación de registros demográficos.

El *Cambridge Group for the History of Population and Social Structure* corona el trabajo de unos 25 años —que comienza cerca de 1970—, sobre la historia de la población inglesa, con dos libros fundamentales: *Population history of England* (ver [40]) y *English population history from family reconstitution* (ver [39]).

El proyecto de investigación toma como fuente los registros de 26 parroquias anglicanas, que consideran representativas de la situación demográfica no sólo de las parroquias en sí, sino de todo el país durante la época en cuestión. En particular, en el segundo libro el enfoque está centrado en la técnica de reconstrucción de familias, y en mostrar su utilidad para obtener datos precisos y complejos sobre fenómenos de fertilidad, mortalidad y nupcialidad. El trabajo de reconstrucción fue realizado por voluntarios, que generaron un total de 530 tablas con información agregada sobre las poblaciones en estudio.

Además del estudio de la población inglesa, E. A. Wrigley produce una compilación general sobre la identificación de individuos y la reconstrucción de familias (ver [41]), uno de cuyos artículos puede considerarse punto de partida del presente trabajo: *Nominal record linkage by computer and the logic of family reconstitution*, del mismo Wrigley y R. S. Schofield.

Unos pocos años después de la primer conferencia en Princeton, y de la publicación de este trabajo de Wrigley, se organiza la *Conference on Methods of Automatic Family Reconstitution*, en 1978, con el objetivo principal de discutir *"alternative methods of family reconstitution"* (ver [34]). Aquí se analizan principalmente las primeras técnicas informáticas aplicadas a la reconstrucción de familias de manera automática, con participación principalmente de investigadores no ingleses (del resto de Europa, Estados Unidos y Canadá).

Avanzando en el tiempo, cabe destacar el importante trabajo, en Francia, de Marion Selz-Laurière y sus intentos por sistematizar, formalizar, automatizar e incluso aplicar técnicas de Inteligencia Artificial a la reconstrucción de familias (ver [27, 28, 29, 30, 31, 32, 33]).

Las corrientes cuantificadoras en la historia, de las cuales ya hemos hablado, más el creciente interés en la demografía y la reconstrucción de familias, junto con un alto grado de optimismo y excitación frente a las posibilidades abiertas por la informática, llevan a muchos en los años '70 a creer en la posibilidad de una reconstrucción 100% automática (cabe recordar, por esa misma época, la gran excitación existente en el campo alrededor de la Inteligencia Artificial). Nacen así algunos paquetes de software que persiguen tal objetivo.

Pronto pudieron verse los límites de estos paquetes y de la posibilidad de reconstruir automáticamente. En los últimos años (década del '90), se produjo un fenómeno inverso. Muchos historiadores se convencieron de la imposibilidad de la reconstrucción masiva de familias, decidieron que el único camino era

manual, y encararon otras direcciones.

1.4 El proyecto de Sacaca y Acasio

El presente trabajo forma parte de un proyecto para la reconstrucción de genealogías de los pueblos de Sacaca y Acasio, en la Bolivia actual, entre 1690 y 1810.

Frente al panorama antes mencionado, tomamos una posición quizás intermedia: la reconstrucción completa y 100% automática no es posible, pero eso no implica abandonar enteramente cualquier objetivo de reconstrucción. Creemos en la posibilidad de la construcción de genealogías *parciales*, fuertemente apoyadas por procedimientos informáticos, pero en ningún momento dejando de lado la intervención y el criterio del historiador. Luego de presentar el contexto histórico en el cual se enmarca nuestro proyecto, presentaremos la solución informática adoptada, que busca mantenerse dentro del criterio recién mencionado.

1.4.1 Antecedentes

El proyecto de Sacaca y Acasio tiene por objeto analizar, en términos de parentesco, las estrategias matrimoniales prevalecientes entre los indígenas del Alto Perú bajo la dominación española. Un tipo de estudio sumamente en boga en la historia social europea de las últimas décadas, pero que pocas veces ha sido encarado por los historiadores latinoamericanos.

En el caso de los Andes, sin embargo, se pueden citar dos estudios pioneros: el de R. Tom Zuidema sobre el sistema de parentesco incaico (ver [42, 43]) y el de John Earls sobre las categorías parentales entre los indígenas peruanos contemporáneos.

Ambos autores establecieron simultáneamente la existencia, en las sociedades andinas, de prohibiciones matrimoniales entre miembros de una línea masculina y una femenina que comparten un ancestro común hasta la cuarta generación (sin contar al ancestro común como la primera, ver figura 1.1), la cual se tornaba preferencial. Zuidema comienza su estudio con la interpretación de un dibujo de Bocanegra, incluído en un tratado pastoral de 1631 (ver figura 1), y lo continúa con un complejo análisis lingüístico de la terminología de parentesco en el mundo andino. Earls, por el otro lado, basa su hipótesis en observaciones etnográficas realizadas en dos comunidades del Perú actual. Los dos autores también destacaron la coexistencia de estas prohibiciones con las estrategias preferenciales de alianza entre ramas colaterales cada dos generaciones. Más aún, Earls plantea un modelo sobre cómo estas prohibiciones matrimoniales pueden combinarse con las alianzas preferenciales dentro de comunidades pequeñas.

Aunque reconociendo el interés de los trabajos de Zuidema y Earls para el análisis general del parentesco en el mundo andino, Françoise Héritier marcó con gran precisión los límites de los mismos. Se trata de trabajos que no están basados en fuentes con peso estadístico, sino en teorías contradictorias soportadas por evidencia textual de difícil interpretación (en el caso de los incas), o en un número limitado de observaciones (en el caso de Earls y sus comunidades contemporáneas); ver [13, 14].

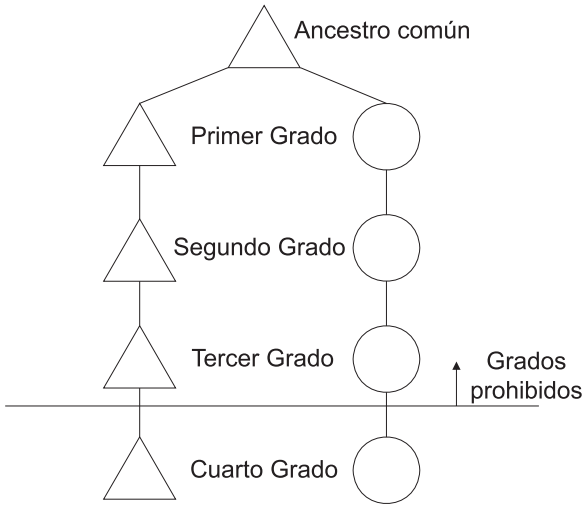

Figura 1.1: Cálculo del grado de consanguinidad a partir de un ancestro común. Earls y Zuidema establecen la existencia de prohibiciones matrimoniales hasta el tercer grado.

El proyecto actual retoma estas críticas, y se plantea como objetivo analizar las estrategias matrimoniales sobre una base estadísticamente significativa, y a lo largo de un período de tiempo considerable dentro del período colonial (siglo XVII-XIX), ubicado temporalmente en el medio entre el período incaico de Zuidema y el contemporáneo de Earls.

Para producir dicha fuente estadísticamente significativa, el objetivo es establecer genealogías, lo más largas posibles en el tiempo, que revelen la mayor cantidad de vínculos posibles, tanto de afinidad como de consanguinidad, entre los contrayentes matrimoniales.

La tarea que describiremos en este trabajo es la reconstrucción de dichas genealogías, proponiendo una metodología general para la reconstitución automática parcial de familias.

La fuente utilizada son las 11750 actas matrimoniales, digitalizadas en el marco de este proyecto, que cubren el período 1690 - 1810 en los pueblos de Sacaca y Acasio.

1.4.2 Contexto

Antes de comenzar la descripción de las herramientas y metodología desarrolladas, comentaremos brevemente sobre el contexto histórico del proyecto.

Recordemos que en el adoctrinamiento de los indígenas a la Fe Católica que siguió a la conquista de América, el matrimonio jugó un rol central (ver [38]). La doctrina de la Iglesia en Europa prohibía matrimonios entre parientes cercanos, dentro del cuarto grado de consanguineidad o afinidad (ver [9]). Sin embargo, la gran discrepancia entre las prácticas observadas entre los indígenas de México y las prohibiciones de la Iglesia llevaron al Papa Pablo III, ya en 1537, a otorgar a los nativos americanos, en su *Bula Altitudo divini consilii*,

una reducción de la prohibición del cuarto al segundo grado, para favorecer su conversión al cristianismo (ver [26]). En los Andes coloniales, tanto en la documentación colonial en general, como en la literatura pastoral de la época, se puede observar una actitud permisiva por parte de la Iglesia Católica al enfrentarse con matrimonios indígenas de segundo grado.

Por ende, creemos que es importante explorar las prácticas y los significados relativos a las estrategias matrimoniales indígenas. El proyecto se enfoca sobre estos dos pueblos de la región de Chayanta en la actual Bolivia: San Luis de Sacaca, en la puna, y San Juan de Acasio, su *anexo* colonial en el valle. Esta región presenta varios caracteres distintivos durante su período colonial. El primero de ellos es su vecindad con el gran centro minero de Potosí, donde el control colonial era aplicado con máxima intensidad (ver [36]). Por lo tanto, las cargas coloniales como el tributo y la mita (la migración indígena forzada hacia el trabajo en las minas de plata), se sentían especialmente en Chayanta durante el período. Sin embargo, esa misma vecindad también estimulaba la especialización de las comunidades indígenas en la producción de abastos para los mercados urbanos, sobre todo de trigo (ver [37]). Esto explica que, en el largo plazo, Chayanta se mantuviera como un caso particularmente exitoso dada la continuidad de su control comunal sobre las tierras de diferentes capacidades ecológicas, tanto en la puna como abajo en los valles, principalmente a través de migraciones estacionales (ver [24]).

1.5 Nuestro trabajo

Como hemos mencionado antes, existieron numerosos proyectos exitosos de construcción de genealogías en el pasado, aunque prácticamente en todos los casos involucrando poblaciones europeas o de América del Norte (tanto en Estados Unidos como en Canadá, poblaciones con nombres ingleses o franceses).

No contamos con antecedentes exitosos de reconstrucciones masivas basadas en fuentes de poblaciones de América Latina, no importa el período de tiempo.

El primer desafío fue, entonces, adaptar las experiencias exitosas, basadas en otras poblaciones, a la población indígena de los Andes coloniales. Varias limitaciones fueron detectadas de forma temprana. Por ejemplo, los métodos existentes para la unificación de variaciones ortográficas en los nombres de los individuos, diseñados pensando en apellidos ingleses, franceses, holandeses, etc., resultan de poca utilidad al ser aplicados sobre este tipo de población (ejemplos de estos métodos son el método SOUNDEX y algunas de sus variantes).

En segundo lugar, los distintos métodos y herramientas existentes fueron en general diseñados para usos y proyectos específicos, y pensados para un tipo de fuente en particular.

Nuestro trabajo se plantea entonces distintos objetivos:

- En primer lugar, reconstruir las genealogías de Sacaca y Acasio.
- Al mismo tiempo, proponer una metodología general para cualquier proyecto de reconstrucción de familias, organizándolo en etapas y pasos bien identificados. El objetivo es que dicha metodología resulte aplicable a cualquier tipo de fuente, referida a cualquier tipo de población, en cualquier período histórico.

- Finalmente, proveer una herramienta computacional de uso general que sustente la metodología planteada.

Comenzaremos por introducir las dificultades inherentes a la identificación de individuos. Luego, plantearemos la estructura de nuestra solución propuesta.

1.5.1 Identificando individuos

La reconstrucción de familias se realiza mediante la identificación de individuos. Identificar implica descubrir cuándo dos instancias de una o varias fuentes distintas hacen referencia a un mismo individuo real.

Se reconstruye demográficamente una vida detectando el acta de nacimiento, matrimonio y defunción del mismo individuo en cuestión. Puede ser de interés también relacionar un registro censal con otro para detectar migraciones, o vincular un registro de votación nominal con actas de propiedad, para analizar el comportamiento político de un sector social.

La reconstrucción de familias en particular intenta reconstruir el árbol genealógico de una familia, identificando por ejemplo dos actas de nacimiento diferentes con los mismos padres, y luego identificando al niño nacido como padre en un acta de nacimiento futura.

En el caso de Sacaca y Acasio, donde la reconstrucción se realiza a partir de actas matrimoniales solamente, se intenta una triple identificación:

- La reaparición de un contrayente como padre en el casamiento de sus hijos, para extender la línea verticalmente, agregando una generación.
- La reaparición de un individuo como padre en dos actas diferentes, extendiendo horizontalmente la línea al detectar hermanos.
- La reaparición de un contrayente casándose nuevamente.

La tarea de identificación de individuos, en los términos antes mencionados, se topa con dos tipos de problemas básicos:

- Las variaciones ortográficas, o de uso en los nombres —sobre todo en poblaciones y con documentos del pasado—.
- La existencia de nombres repetidos dentro de una población. ¿Cuántas *María Mamani* nacen, se casan, o mueren un mismo año en el Altiplano boliviano?

Si los individuos estuvieran registrados en las fuentes mediante identificadores o claves únicas —o prácticamente—, todo este problema no existiría. En la actualidad, muchas burocracias estatales utilizan en sus registros números de documento, de seguridad social, o similar, para identificar unívocamente al individuo en cuestión.

En documentos del pasado, los individuos están generalmente identificados por su nombre y apellido, y algunos datos anexos tales como ocupación, fecha de nacimiento, lugar de residencia, nombre de sus padres, etc.

En este tipo de documentos, si los individuos involucrados son hombres ricos, o poderosos, o relacionados a hechos históricos de relevancia, las ambiguedades son fácilmente resueltas con un mínimo de conocimiento del contexto, que cualquier historiador embebido en la época y el lugar debe poseer.

Sin embargo, cuando todas nuestras fuentes se refieren a este *hombre común*, y cuando la tarea de identificación se realiza masivamente sobre toda una fuente, los posibles errores de identificación deben atacarse de otra forma.

Las variaciones ortográficas pueden darse por diversas situaciones:

- En muchos casos, tratándose de fuentes del pasado, la ortografía de un nombre depende de quien lo registra. El párroco que casa a *Diego Hachayo* en 1712 puede ser distinto al que casa a su hijo *Jose Achallo* en 1740. También ocurre que la ortografía "correcta" de un nombre o un apellido varía con el tiempo.

- En otros casos, varía la forma de registrar al mismo individuo. No siempre se utiliza su nombre completo, a veces se abrevia el nombre de pila, etc. ¿Es el *Juan Miguel Díaz de Solís* que compra esta tierra en 1678 el mismo *J. Solís* que muere en una parroquia cercana en 1690? ¿Será su padre?

- Por último, pueden producirse errores de tipeo o interpretación al digitalizar la fuente. Es común que este trabajo lo realice, por ejemplo, un dataentrista a partir de un acta manuscrita microfilmada, de difícil lectura.

En la figura 1.2 puede verse la versión microfilmada de un acta original. En primer lugar, se hace obvio el problema de la legibilidad. En segundo lugar, el ejemplo muestra un caso interesante de variación ortográfica. Leyendo con cuidado, puede verse que el novio es de apellido `Collque`, mientras que su padre está registrado como `Colque`, con `l` en lugar de `ll`.

Figura 1.2: Parte de un acta matrimonial, tal como es visualizada a través del microfilm.

En resumen, se hace claro que existen dos requisitos básicos para identificar reapariciones de individuos.

El primero, estandarizar de alguna forma la ortografía. Aquí se corre un doble riesgo: si dos variaciones ortográficas no se unifican, se pierde luego cualquier

posibilidad de vinculación; al mismo tiempo, si se unifican incorrectamente se generan vinculaciones falsas.

El segundo requisito es definir un conjunto de reglas que permitan decidir en forma automática cuándo dos apariciones de nombres estandarizados, que coinciden total o parcialmente, hacen referencia al mismo individuo real. En este caso, nuevamente, se corre un riesgo doble: reglas demasiado estrictas dejarán afuera muchas posibles identificaciones, dando resultados pobres, mientras que reglas muy laxas generarán muchas falsas identificaciones.

Comentaremos a continuación la estructura general que proponemos para un proyecto de reconstrucción de esta naturaleza. En el próximo capítulo desarrollaremos detalladamente la forma en que aplicamos la metodología a las actas de Sacaca y Acasio, identificando los problemas encontrados, las soluciones propuestas y las herramientas desarrolladas.

1.5.2 Etapas de nuestra solución

Desechada la posibilidad de la reconstrucción de familias de forma totalmente automática, nuestra solución consistió en establecer una serie de etapas bien diferenciadas para la construcción de genealogías parciales, cada una de ellas sostenida por una o varias herramientas informáticas.

Sostenemos que estas etapas son comunes a todo proceso de reconstrucción genealógica y que, junto con las herramientas provistas, pueden ser llevadas adelante en cualquier proyecto de este tipo, independientemente de la época, la lengua o la cultura involucradas.

Nuestra experiencia inicial con algunos intentos de mayor automatización nos mostraron la importancia de establecer etapas claramente divididas y que puedan ser cerradas ordenadamente una a una.

Las etapas involucradas son:

Etapa 1 - Estandarización de nombres: Todo proceso debe comenzar por algún tipo de homogeneización ortográfica entre los nombres involucrados.

Etapa 2 - Estandarización de datos anexos: Estandarizados los nombres, se debe categorizar, codificar o estandarizar otros datos asociados que puedan servir a la identificación, tales como ocupaciones, lugares de nacimiento o residencia, estados civiles, etc.

Etapa 3 - Vinculación de registros: Esta es la etapa de identificación propiamente dicha, donde se establecen y se *corren* los criterios de identificación de individuos.

Etapa 4 - Detección y eliminación de inconsistencias: El proceso de vinculación de registros puede conducir a identificaciones inconsistentes, que deben ser resueltas y eliminadas.

En el **capítulo 2** examinaremos métodos existentes de estandarización de nombres, principalmente *Soundex* y sus derivados, que fueron testeados con pobres resultados. Identificamos la debilidad y poca adaptabilidad de los métodos, y proponemos una solución propia de estandarización basada en reglas de equivalencias ortográficas, mostrando cómo se utilizó esta solución en el caso de Sacaca y Acasio (**etapa 1**). Veremos que la solución propuesta permite al

historiados codificar su propio conocimiento contextual y de la lengua en cuestión. A su vez, destacamos nuevamente la insuficiencia del método automático, y la necesidad de intervención del historiador para revisar y reagrupar manualmente ciertos términos unificados. En el caso de Sacaca y Acasio, se convocó a lingüistas especializados para este proceso.

En el mismo capítulo revisaremos la **etapa 2**, de estandarización de datos asociados, tales como la ocupación de un individuo, referencias geográficas, etc. En primer lugar, pueden existir simples variaciones ortográficas en estos datos. Pero, en segundo lugar, pueden existir formas diferentes de referirse a una misma profesión, categoría fiscal o lugar geográfico, y los datos pueden aparecer con distinto grado de agregación (por ej., para un mismo individuo puede hacerse referencia a su lugar de nacimiento a través del nombre del pueblo, la región, la provincia o el país). Aquí también la experiencia del historiador es esencial y veremos distintas formas de parametrizar las herramientas en función de este conocimiento.

En el **capítulo 3** describiremos detalladamente el problema de identificación de individuos mediante *vinculación de registros* (**etapa 3**) y las herramientas desarrolladas a tal fin. Veremos cuáles son los datos a vincular, la forma de validar y *pesar* las vinculaciones, y la forma en que un conjunto de identificaciones nos permite construir una genealogía. Analizaremos luego las posibles inconsistencias que el proceso genera y su posible solución (**etapa 4**).

En el **capítulo 4** definiremos algunas métricas para la eficiencia y el éxito del proceso de reconstrucción de genealogías históricas. Por último, analizaremos el proyecto de Sacaca y Acasio de acuerdo a estas métricas.

El **capítulo 5** contiene las conclusiones, algunas menciones al trabajo actual que estamos realizando, y propuestas para trabajo futuro.

Capítulo 2

Normalización de nombres y términos asociados

2.1 El problema

Cualquier proyecto de *record linkage* entre fuentes históricas debe comenzar por algún tipo de homogeneización ortográfica. Más allá de los criterios a adoptar luego en la identificación efectiva de los individuos, el punto de partida del proceso necesariamente será algún tipo de coincidencia nominal entre personas, y para que dicha coincidencia exista debe sobrepasarse la barrera de las variaciones ortográficas.

Repasemos las razones, mencionadas en la Introducción, por las que se producen estas variaciones.

En primer lugar, no debemos descartar el error propio de la digitalización. Ya se trate de un proceso de transcripción manual de la fuente, o de digitalización automática vía OCR, necesariamente se producirán errores. En el caso de la transcripción manual, los errores pueden ser tanto de tipeo como de interpretación. En nuestro caso, basta observar alguna de las actas manuscritas —en su versión microfilmada sobre todo— para comprender la dimensión del problema (ver figura en Introducción). Distintos dataentristas involucrados probablemente interpreten de forma diferente un mismo grafismo. A esto debemos agregar los cambios introducidos por el escriba original en sucesivos registros, o por los distintos escribas que copian o transcriben la fuente.

En segundo lugar, en la mayoría de los casos el nombre de un individuo es enunciado oralmente por él mismo y registrado por un funcionario, de quien dependerá entonces la ortografía elegida. Debemos tomar en cuenta que en muchos casos, además, se trata de poblaciones con alto grado de analfabetismo, donde por más que el individuo tuviera oportunidad de validar el registro, no está capacitado para hacerlo. Es así como la ortografía depende mayormente de la opinión del registrante, sea este un párroco, un recaudador de impuestos, un censista o cualquier otro agente. Las migraciones, la diversidad de lenguas, la multiplicidad de nombres, impiden el consenso ortográfico total.

En tercer lugar, la ortografía no es estática, y evoluciona con el tiempo. Podemos citar el ejemplo del apellido inglés *Smythe* que evoluciona hacia el más actual *Smith*.

No tomaremos en cuenta en esta sección el caso de los diferentes usos de un nombre (en la introducción mencionamos el ejemplo de un *Juan Díaz de Solís* contra un *J. Solís* en dos apariciones distintas en una fuente), ya que se trata de un problema diferente al de la variación ortográfica. Este problema debe ser atacado en la etapa de record linkage propiamente dicha.

2.2 Soluciones existentes: el método Soundex

El primer método desarrollado para la estandarización de nombres, y aún hoy el más popular, es el método *Soundex*.

El método fue desarrollado por Margaret K. Odell y Robert C. Russell en 1918 para el US Bureau of Archives, con el objetivo de simplificar la registración y recuperación de información censal.

Aunque existen numerosas variantes, la idea básica consiste en codificar cada nombre de acuerdo a su sonoridad. Se parte de la base de que es esperable que variaciones ortográficas de un mismo nombre se pronuncien de manera muy similar, y por ende se traduzcan al mismo código Soundex.

Soundex codifica los nombres en base a la siguiente tabla:

Letra	Codificación
B,F,P,V	1
C,G,J,K,Q,S,X,Z	2
D,T	3
L	4
M,N	5
R	6
H,Y,W	se omite
A,E,I,O,U	se omite

En su forma canónica (ver [17]) —ya mencionamos que existen variantes—, la codificación se realiza de la siguiente forma: se mantiene el primer caracter del nombre, y se eliminan todas las vocales no iniciales y todas las H, Y y W no iniciales. Luego se codifica el resto de los caracteres de acuerdo a la tabla anterior, manteniendo sólo un dígito para caracteres consecutivos con la misma codificación (por ejemplo, `CK` codifica a *2*, no a *22*). Finalmente, se trunca la codificación conservando sólo los 3 primeros dígitos, y los nombres que codifican a menos de 3 dígitos son completados con ceros. Veamos algunos ejemplos:

Nombre	Código Soundex
`Smith` y `Smythe`	*S530*
`Gardner`, `Gardiner` y `Gartner`	*G635*
`Duke` y `Diuk`	*D200*
`Lope` y `Leiva`	*L100*
`Lopez`	*L120*
`Hachaya`	*H200*
`Achaia`	*A200*

De los ejemplos anteriores podemos detectar inmediatamente algunas de las primeras limitaciones de Soundex.

En los tres primeros casos (`Smith`, `Gardner`, `Duke`), donde se trata de nombres ingleses, parece comportarse bastante bien y en la forma esperada. Sin

embargo, luego podemos ver que el método identifica incorrectamente a `Lope` con `Leiva`, mientras que no une a `Lope` con `Lopez`. De la misma forma, se pierde de unificar a `Hachaya` con `Achaia`.

A partir de estas observaciones, podríamos intentar modificar el algoritmo variando las asociaciones de consonantes con los números que las codifican.

Observamos por ejemplo la gran cantidad de letras asociadas en el segundo grupo (codificadas a *2*). Esto se produce transitivamente, ya que la letra C debe relacionarse al mismo tiempo con la S y la K, y éstas a su vez con la Z y la Q. El resultado entonces es que Z y Q codifican de la misma forma, cuando es obvia su escasa relación fonética.

Se han propuesto numerosas variaciones al método a partir de estas observaciones: diversas formas de asociar las consonantes; la extensión del código resultante para permitir más de 3 dígitos; la codificación conjunta de ciertas "consonantes compuestas" (por ejemplo, *TCH*, *CH*, etc.); y la codificación de la letra inicial de la misma forma que las otras.

Sin embargo, todas las versiones de Soundex intentan capturar equivalencias fonéticas sin tener suficientemente en cuenta el contexto, dentro de la palabra, de cada letra o conjunto de letras.

Podemos mencionar dos intentos de superar también esta limitación: el método Phonex y el método Daitch-Mokotoff Soundex, aunque ambos fallan en abarcar suficientemente la complejidad fonética-ortográfica de las diversas lenguas y culturas.

Podemos citar el trabajo de Alan Stanier para ver que las limitaciones persisten aún en las variantes (ver [35]). Stanier toma el censo de los Estados Unidos de 1851 y, a partir del *Dictionary of Surnames* de Hanks y Hodges [11], busca cuáles de las variaciones ortográficas de cada nombre reconocidas por dicho Diccionario son correctamente reconocidas por Soundex en cuatro de sus variantes. En todos los casos, no más de un tercio de las variaciones ortográficas que Soundex reconoce son correctas, y aproximadamente un cuarto de las variaciones correctas son pasadas por alto por el método.

Podemos resumir brevemente las limitaciones detectadas en el método Soundex, siguiendo en parte a Patman y Shaefer (ver [23]):

Dependencia de la letra inicial: Esta limitación es muy clara. Se pierde toda variación ortográfica, por mayor coincidencia fonética que haya, sólo por el hecho de que la primer letra debe ser respetada. Dos nombres —extraídos de nuestra fuente—, como `Valdivieso` y `Baldivieso` reciben una codificación diferente.

Falta de adaptación a culturas e idiomas diferentes: Soundex fue diseñado originalmente pensando en nombres ingleses, y aunque existen adaptaciones a otros idiomas, en ningún caso se logra un método general apto para cualquier contexto cultural. Al trabajar sobre bases de nombres de orígenes diversos, varios aspectos deben ser tenidos en cuenta de forma simultánea.

En primer lugar, cuando se trata de nombres provenientes de culturas que utilizan alfabetos no románicos, se deben tener en cuenta las distintas formas de transcripción a nuestro alfabeto. Por ejemplo, se han detectado al menos 12 formas distintas en que se ha transcripto desde el alfabeto cirílico, en distintas ediciones, el nombre del autor ruso `Fyodor Dostoyevsky`

(o `Fyodor Dostoevsky`, `Fedor Dostoievsky`, `Fjodor Dostojewskij`, `Fedor Dostojewski` , etc.).

En segundo lugar, en distintos idiomas existen diversos usos de consonantes mudas. Un caso es el del inglés, donde Soundex no puede manejar las pronunciaciones similares de nombres como `Deighton` y `Dayton`, o `Coghburn` y `Coburn`.

En tercer lugar, algunas culturas utilizan en sus nombres prefijos o elementos opcionales. Tomemos un ejemplo del árabe, donde nombres como `Al-hameed` pueden a veces prescindir de su prefijo transformándose en `Hameed` (y sus variantes `Hamid`, `Hamed`, etc.).

Por último, se debe tener en cuenta la equivalencia de nombres o apodos. Es común encontrar citas a, por ejemplo, una `Margaret Jones` como `Peggy` o `Maggy Jones`.

Poca tolerancia a errores de tipeo: Errores típicos de tipeo, como la omisión o la inversión de caracteres, pueden producir fácilmente codificaciones distintas: `Leanro` (*L560*) en lugar de `Leandro` (*L536*), o `Avlarado` (*A146*) por `Alvarado` (*A416*).

2.3 La solución implementada: un sistema basado en reglas

A partir del análisis del método Soundex y sus variantes, surge la necesidad de diseñar una metodología y una herramienta general para la codificación de los nombres.

Siguiendo a Morris en [22]: *"[We] need to develop other systems according to the cultural and language base of the records concerned".*

La base de nombres de Sacaca y Acasio se constituye a partir del intercambio y mestizaje entre aymaras y españoles, con la riqueza y diversidad que ello implica. Antes de cualquier proceso de homogeneización, la base está constituída por más de 6000 nombres y apellidos distintos.

2.3.1 Definición y aplicación de reglas

Diccionario de reglas de traducción

Se desarolló entonces una herramienta que permite al historiador definir su propia base de equivalencias ortográficas, o semánticas, a partir de su conocimiento del contexto y el idioma en estudio.

El historiador construye un diccionario de equivalencias entre cadenas de caracteres, tomando en cuenta no sólo la cadena en sí sino su posición dentro de un término.

Por ejemplo, comienza por definir que la letra V es equivalente a la letra B, no importa en qué posición aparezca, o definir que BE como terminación de un nombre o apellido es equivalente a PE.

Notaremos estos dos casos de la siguiente manera:

$$\alpha_1 V \alpha_2 \equiv \alpha_1 B \alpha_2,$$
$$\alpha_1 PE \equiv \alpha_1 BE,$$

donde α_i representa una cadena cualquiera, incluyendo la cadena vacía λ.

La herramienta permite, en el proceso de definición de estas reglas, consultar todos los términos que se verían afectados en cada caso. Esto permite al historiador verificar y descubrir tempranamente posibles casos de asociaciones erróneas, o nuevas posibles asociaciones para construir su base de homogeneización ortográfica.

El siguiente paso en el proceso es transformar las reglas de equivalencia en traducciones, de la forma:

$$s_1|..|s_n \rightarrow cod$$

Esto significa que las cadenas s_1 a s_n, dentro del término, serán codificadas (traducidas) a la cadena *cod*. Un diccionario es, por lo tanto, una gramática extendida por la posibilidad de definir las posiciones dentro de un término.

En el caso de las dos reglas planteadas como ejemplo, el historiador debe decidir cuál de las cadenas equivalentes será la cadena codificadora, la cadena a la cual se traducirán las demás. El resultado puede ser, por ejemplo:

$$\alpha_1 V \alpha_2 \rightarrow \alpha_1 B \alpha_2 \text{ ó } \alpha_1 B \alpha_2 \rightarrow \alpha_1 V \alpha_2,$$
$$\alpha_1 PE \rightarrow \alpha_1 BE \text{ ó } \alpha_1 BE \rightarrow \alpha_1 PE.$$

Las reglas se definen para las siguientes posiciones posibles (llamemos S a la cadena a codificar):

$S\alpha$	Al inicio del término.
αS	Al final del término.
$\alpha S \gamma$	En cualquier ubicación dentro del término.
_ $\alpha S \gamma$	En cualquier ubicación dentro del término menos al comienzo.
$\alpha S \gamma$ _	En cualquier ubicación dentro del término menos al final.
_ $\alpha S \gamma$ _	En cualquier ubicación dentro del término menos al comienzo o al final.
S	La cadena debe ser el término completo.

El diccionario utilizado para Sacaca y Acasio está compuesto por 141 reglas diferentes de este tipo. Veamos un par de ejemplos tomados del mismo:

$$GUA\alpha \rightarrow HUA\alpha$$
$$\alpha H \gamma \rightarrow \alpha \gamma$$
$$\alpha PA \rightarrow \alpha BA$$
$$\alpha B \gamma \rightarrow \alpha V \gamma$$
$$FRANCISCA \rightarrow CISCA$$

Antes de ver cómo se aplica este diccionario al universo de nombres y apellidos, y estudiar algunas limitaciones del algoritmo, veamos de qué manera el diccionario permitiría superar algunas de las fallas mencionadas en el análisis de Soundex:

Transcripción desde otros alfabetos: Es posible establecer reglas de equivalencia para variantes de transcripción conocidas. Por ejemplo, la regla $FE\alpha|FY\alpha \rightarrow FI\alpha$, o directamente `DOSTOIEVSKY` → `DOSTOEVSKI`.

Consonantes mudas: Habría que definir las reglas correspondientes, como $\alpha GH\gamma \rightarrow \alpha\gamma$.

Uso de prefijos: Podríamos definir la regla $AL\text{-}\alpha \rightarrow \alpha$, para el caso mencionado de `Al-Hameed`.

Equivalencia de nombres o apodos: La regla $MARGARET|MAGGIE|MAGGY \rightarrow PEGGY$ resuelve el problema.

Definición de las reglas

La definición del conjunto de reglas se realiza a partir de distintas fuentes: el conocimiento contextual del historiador, el aporte de lingüistas especializados en el idioma en cuestión, la experiencia con la fuente, la percepción y experiencia de los dataentristas durante la digitalización, y por supuesto el mero sentido común.

Durante el proceso de definición de reglas, hemos utilizado el método Soundex como un agrupador inicial. A partir de los errores y aciertos del método, se facilita la tarea de detectar reglas que agrupen o separen variaciones ortográficas.

Al mismo tiempo, como hemos mencionado antes, la herramienta permite en todo momento consultar el conjunto de términos afectados por cada regla, junto con la forma en que dicha traducción generaría equivalencias.

Procesamiento y generación de las variaciones ortográficas

Una vez definido el diccionario de reglas de traducción, se ejecuta el proceso de generación de variaciones ortográficas. El mismo comienza tomando el conjunto de nombres y apellidos de la base como el *corpus* inicial.

El siguiente paso consiste en crear, para cada uno de los términos del *corpus*, un nuevo conjunto que lo incluya como único elemento, identificándolo dentro de dicho conjunto como *término original*.

Utilizaremos como ejemplo del proceso dos términos de nuestra base de Sacaca y Acasio, los apellidos `HUANPA` y `GUAMBA`. El proceso comenzará entonces partiendo de los conjuntos:

{HUANPA}
{GUAMBA}

Para cada uno de los conjuntos de términos, el proceso recorrerá repetidas veces el conjunto de reglas de traducción. Al encontrar alguna de las cadenas izquierdas de una regla, en la posición correcta, el proceso generará un nuevo término igual al original, al que le aplicará la traducción correspondiente. Este nuevo término se incorpora a su vez al conjunto, para ser eventualmente afectado por otra nueva regla de traducción.

Veamos un ejemplo con nuestros dos conjuntos iniciales, y algunas reglas que los afectan:

> La regla $\alpha PA \rightarrow \alpha BA$ afectará al término *HUANPA*. La traducción impone la generación de un nuevo término: *HUANBA*. Dicho término es incorporado al conjunto, resultando ahora en:

{HUANPA, HUANBA }

Una segunda regla nos dice que la N no inicial puede traducirse como M: $\_\alpha N\gamma \rightarrow \_\alpha M\gamma$. Esto volverá a afectar a nuestro conjunto, que resulta ahora:
{HUANPA, HUANBA, HUAMPA, HUAMBA}

Por otro lado, una tercer regla de traducción nos dice que $GUA\alpha \rightarrow HUA\alpha$. Esta regla afectará al segundo conjunto original de términos, el conjunto que inicialmente incluye solamente a *GUAMBA*. Luego de aplicar esta regla, el conjunto se transforma en:
{GUAMBA, HUAMBA }

El proceso se repite con todas las reglas y, una vez aplicadas, vuelve a comenzar desde la regla número uno. De esta manera, la base de términos crece en cada iteración. En nuestro caso anterior, aplicando repetidamente todas las reglas, podemos ver que nuestros dos conjuntos de ejemplo se transforman en conjuntos de más de 80 variaciones cada uno, incluyendo elementos como:

{GUAMPA, GUAMBA, GUAMVA, GVAMBA, GBAMBA, BAMPA, BAMBA, VAMBA, UAMBA, UAMVA, ...}

El proceso así definido generará, a partir de las reglas de traducción, conjuntos que incluyen todas las variaciones ortográficas posibles de un término.

El último paso consiste entonces en identificar intersecciones entre estos conjuntos, que inicialmente eran disjuntos. Si dos conjuntos intersecan, definiremos que los dos términos que originalmente los constituían (y que fueron marcados oportunamente) constituyen una variación ortográfica de un mismo nombre dentro de la base.

Los conjuntos {*HUANPA, HUANBA, HUAMPA, HUAMBA*} y {*GUAMBA, HUAMBA*} intersecan a través del término $HUAMBA$, por lo que consideraremos que los elementos originales $HUANPA$ y $GUAMBA$ son variaciones del mismo nombre.

El primer obstáculo con el que nos tropezamos en esta forma de procesamiento es el crecimiento eventualmente exponencial —y eventualmente infinito— de cada conjunto de términos. Cada nueva regla que se incorpora al diccionario podría eventualmente afectar a todos los términos de un conjunto inicial, generando que se dupliquen los elementos de dicho conjunto.

Llamemos R al conjunto de reglas, C a un conjunto inicial de términos, y $L_C(R)$ al conjunto que determina, a partir de C, la aplicación de todas las reglas de R. Definamos $R' = R \cup \{\alpha \rightarrow \gamma\}$. Podemos ver que en el peor caso, si la regla agregada afecta al conjunto C completo, $\#L_C(R') = 2\#L_C(R)$.

Por otro lado, si el conjunto de traducciones no se define con cierto cuidado, no hay garantías de que $L_C(R)$ sea finito. Un ejemplo trivial: supongamos que existiera una regla de traducción del tipo $\alpha \rightarrow \alpha\alpha$. El conjunto inicial $\{\alpha\}$ se transformaría luego del primer paso en $\{\alpha, \alpha\alpha\}$, en un segundo paso en $\{\alpha, \alpha\alpha, \alpha\alpha\alpha\alpha\}$ y así sucesivamente.

Aunque el caso dado en el ejemplo es elemental, situaciones similares pueden darse por encadenamiento de reglas, de una forma más compleja de detectar.

Más adelante mostraremos cómo generar un grafo dirigido con las reglas de traducción, que permita detectar estas situaciones identificando ciclos en el mismo.

Los dos problemas mencionados: el crecimiento de los conjuntos acotado exponencialmente, y la posible infinitud de los mismos, obliga a definir un criterio de detención del algoritmo:

1. Sobre cada conjunto inicial, el proceso se detiene si luego de una pasada completa por todas las reglas no se agregan términos.

2. Si esto no ocurre, el proceso simplemente se detiene luego de n pasadas.

Con la experiencia hemos visto que, más allá de la quinta pasada, no se generan variaciones ortográficas de interés. Se debe tener en cuenta que la cantidad de pasadas por todo el conjunto de reglas tiene relación directa con la posible cantidad de variaciones ortográficas dentro de un mismo término original que podrían llevarlo a intersecarse con otro término. En nuestro ejemplo de $HUANPA$ y $GUAMBA$, identificamos 3 variaciones: HUA por GUA, N por M y PA por BA. De acuerdo al orden en que se procese cada regla de traducción, la confluencia de estos términos se produce en una o dos pasadas.

2.3.2 Ventajas y problemas del sistema basado en reglas

Los dos problemas básicos mencionados en nuestro proceso basado en reglas, su exponencialidad y su infinitud, hacen pensar en la posible necesidad de recurrir a otro mecanismo de unificación ortográfica.

Una posibilidad consiste en recurrir, en lugar de a reglas de traducción que generan nuevos términos, a un mecanismo de codificación de cadenas, similar al utilizado en *SOUNDEX*.

De acuerdo a este mecanismo, definiríamos reglas del siguiente tipo, para ser aplicadas una vez sola en una sola pasada:

$$\begin{aligned} V \mid B &\rightarrow 1 \\ HUA \mid GUA &\rightarrow 2 \\ PA \mid BA &\rightarrow 3 \end{aligned}$$

Sin embargo, mostraremos algunas situaciones donde este mecanismo no es efectivo:

PROBLEMA DE PRECEDENCIA Supongamos que contamos con las siguientes dos reglas: $\alpha H\gamma \rightarrow \alpha\gamma$ y $\alpha PH\gamma \rightarrow \alpha F\gamma$. ¿Cuál debe aplicarse primero? Si aplicamos la primera de ellas, un nombre como $PHELIPE$ se traduciría en $PELIPE$, y nunca unificaría con $FELIPE$.

TRANSITIVIDAD Tomemos como ejemplo las reglas $\alpha V\gamma \rightarrow \alpha B\gamma$ y $\alpha BE \rightarrow \alpha PE$. En principio, podríamos imaginar que esto implique que la regla $\alpha VE \rightarrow \alpha PE$ también debería aplicar.

En el primer caso, el ejemplo sugiere que la precedencia es fácil de resolver. Todo parece indicar que PH debe ser traducido antes que H. Pero esto obligaría, por un lado, a que el historiador defina claramente las precedencias en los casos

donde sea necesario, lo que ocurre cuando la parte izquierda de una reglas de traducción está incluída en otras parte izquierda. El caso puede no ser tan claro en situaciones de inclusión total. Veamos el siguiente ejemplo:

> La regla $V \rightarrow B$ podría ser de utilidad para unificar un caso como $BACA$ vs. $VACA$.
> En otro caso, la regla $V \rightarrow U$ sería necesaria para unificar $AYAVIRE$ con $AYAUIRE$.

No es claro cómo, en este ejemplo, debería aplicarse un criterio de precedencia.

Frente al segundo problema, cabe discutir si la transitividad es o no deseable. Retornando al ejemplo anterior, imaginemos que las reglas son de la forma $B \rightarrow V$ y $V \rightarrow U$. No suena deseable aplicar una regla transitiva, que implícitamente defina la traducción $B \rightarrow U$. Sin embargo, tomemos el caso $V \rightarrow B$ y la regla $BE \rightarrow PE$. Aquí suena más razonable definir implícitamente que $VE \rightarrow PE$, como se puede ver en los nombres $UARPE$ y $UARVE$ de nuestra base.

Nuestra forma de procesar las reglas determina entonces que la transitividad exista, y que la definición de la precedencia no tenga importancia, ya que eventualmente todos los casos posibles serán generados (ver figura 2.1).

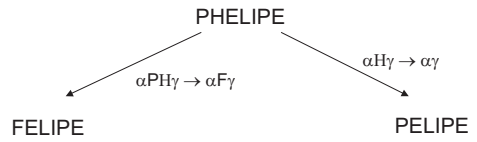

Figura 2.1: Aplicación de reglas a los nombres PHELIPE y FELIPE. No es necesario definir explícitamente precedencias ni transitividades.

El criterio utilizado, que genera grandes cantidades de variaciones ortográficas para cada término, aún algunas que parece carecer de sentido, se basa entonces en dos conclusiones:

- No debemos exigir del historiador extremo cuidado ni completitud al definir las reglas. Es preferible que se generen posibles conjuntos infinitos, interrumpiendo el proceso a las n pasadas, antes que perder unificaciones.
- Los casos de variaciones absurdas difícilmente generen una intersección con otro conjunto. Por otro lado, resulta más fácil, en una segunda etapa, desarmar grandes grupos de términos erróneamente unificados que detectar casos de no unificación.

En conclusión, este proceso de generación masiva de variantes ortográficas ha demostrado ser útil y efectivo a la hora de detectar términos originales que pueden definirse como equivalentes, a riesgo de ser lento y en algunos casos excesivamente generoso con la unificación.

Mejorando la herramienta

Por último, agregaremos que es posible incorporar algunas funcionalidades dentro de la herramienta que ayuden al historiador a depurar y mejorar su diccionario de traducciones.

En primer lugar, es conveniente sugerir que las reglas de traducción se definan colocando en su parte derecha cadenas de longitud menor o igual a las cadenas de la parte izquierda. Esto permitiría que la cantidad de variantes ortográficas para un término converja, que sea finita.

En segundo lugar, como mencionamos antes, es posible construir un grafo que permita detectar encadenamientos de reglas que lleven a una explosión de variantes.

Construimos el grafo de la siguiente manera:

1. Para cada cadena definida en cualquiera de las reglas, creamos un nodo.
2. Por cada regla del tipo $a_1|..|a_n \rightarrow cod$ agregamos n ejes dirigidos, desde los nodos correspondientes a cada a_i hacia el nodo *cod*.
3. Finalmente, por cada par de regla $a_i \rightarrow cod$ y a_j tal que $a_i \subseteq a_j$, agregamos un eje desde a_j hacia *cod*.

La figura 2.2 muestra el grafo resultante para las cadenas:

$$PA \rightarrow BA$$
$$B \rightarrow V$$
$$V \rightarrow U$$
$$GUE \rightarrow HUE$$

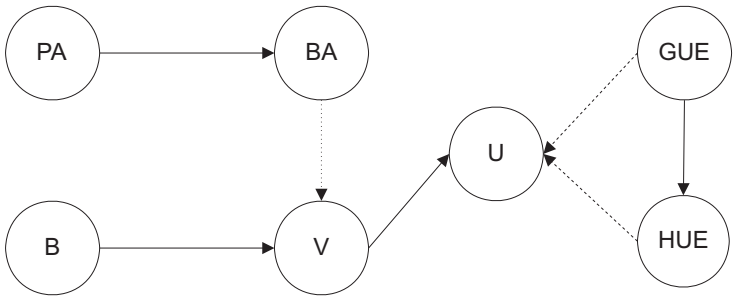

Figura 2.2: Grafo de cadenas. Las líneas punteadas representan ejes generados por el paso 3.

Supongamos que por error, o producto de la gran cantidad de reglas de traducción que intervienen en un proceso de homogeneización ortográfica, existieran además las siguientes reglas [1]:

$$V \rightarrow VE$$
$$VE \rightarrow BE$$

Al agregar estas reglas, el grafo resultante puede verse en la figura 2.3.

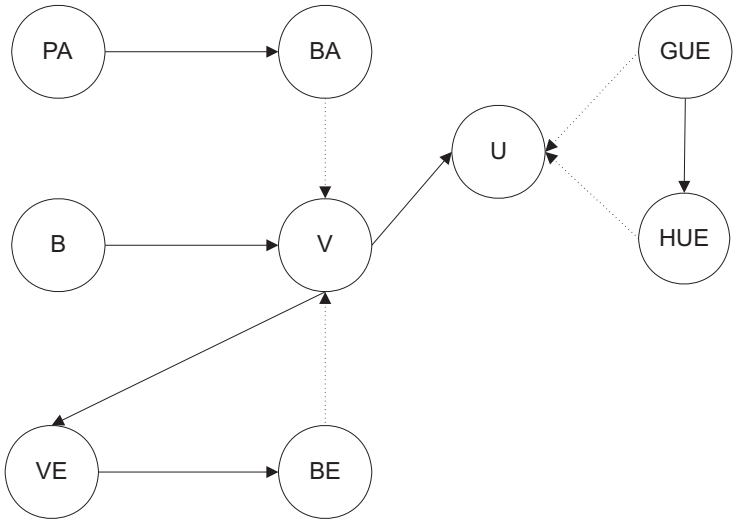

Figura 2.3: Grafo de cadenas, con un ciclo entre los nodos V, BE y VE.

La existencia de este ciclo es la muestra de que un término como $ALAVE$ genere un conjunto infinito de variantes ortográficas. El proceso de generación del grafo de cadenas, y la detección de ciclos dentro del mismo, ha demostrado ser útil para detectar este tipo de errores tempranamente.

2.4 Codificación de términos asociados

Más allá de la homogeneización ortográfica y unificación de nombres y apellidos, el problema se mantiene con otros datos asociados a cada individuo que pueden encontrarse en registros de este tipo.

Algunos datos comunes refieren a profesiones, lugares de residencia o nacimiento, nacionalidades, origen étnico, categoría fiscal, etc.

En estos casos existen, al igual que en los nombres y apellidos, problemas de variación ortográfica. Sin embargo, no es este el mayor de los problemas.

Por un lado, suelen existir distintas formas de referirse a un mismo objeto, sea éste un lugar, un oficio, etc. Los usos y costumbres cambian, y también cambia en algunos casos la granularidad o especificidad con que se nombra una cosa.

Ejemplos claros de esto último se producen tanto con las denominaciones geográficas como con las profesiones. El acta de bautismo de un individuo nacido en un pequeño pueblo, realizada por el párroco del mismo, puede ser bastante detallada en cuanto a su lugar de nacimiento, mencionando como tal hasta una sub-división del pueblo en cuestión. Ese mismo individuo, a efectos fiscales, puede ser identificado simplemente por el condado donde habita o nació. Y si llegara a emigrar a un país lejano, seguramente será simplemente identificado por su nacionalidad, o incluso por el continente de proveniencia.

[1]Cabe aclarar que, aunque el ejemplo parezca no tener demasiado sentido, lo hemos detectado en nuestra base original de reglas para Sacaca y Acasio.

En el caso de las profesiones, un *matricero* puede en otro registro ser identificado como *trabajador metalúrgico*, y en un tercer caso simplemente como *obrero*.

Es importante, entonces, que la herramienta de identificación de individuos permita, además de unificar ciertos términos, establecer categorías de inclusión, que establezcan por ejemplo que un pueblo está dentro de un condado, que a su vez es parte de una provincia, en un determinado país. Puede ser necesario también asociar un pequeño pueblo con alguna ciudad muy cercana, que pueda identificar también la procedencia de un individuo.

En el caso de Sacaca y Acasio, un problema de este estilo lo encontramos con los *ayllus* [2].

Por un lado, observamos una importante disparidad ortográfica en la denominación de los ayllus, que fue unificada manualmente (aunque podría haberse utilizado un proceso como el de los nombres, el número de variantes no lo ameritó).

Por otro lado, existe el caso en que un ayllu determinado se subdivide. Por ejemplo, el ayllu *Chaiquina* se subdividió en *Chaiquina Arriba* y *Chaiquina Abajo*, según si sus miembros estuvieran en la montaña o en el valle.

Por último, todos los ayllus de la zona están agrupados en dos categorías, que llamamos *mitades*.

De esta forma, una jerarquía de inclusión establecería que un individuo identificado como *Chaiquina Arriba* puede en otro caso ser considerado *Chaiquina*, o directamente como miembro de una de las *mitades*. Obviamente, la identificación de los individuos será cada vez más débil, a medida que se realiza asociando categorías más generales.

[2] El ayllu es un subagrupamiento étnico de los indígenas de la zona. El ayllu se transmite de padres a hijos, y una mujer puede ser absorbida por el ayllu de su marido.

Capítulo 3

Vinculación de registros

La etapa principal en el proceso de reconstrucción, una vez homogeneizada la ortografía de los nombres y codificados apropiadamente los datos anexos, es la vinculación efectiva de los registros.

Se trata en este momento de vincular a cada individuo con sus reapariciones en otros registros.

El problema aquí puede ser de dos tipos:

- Aún cuando la ortografía de cada término es homogénea, el nombre completo de un mismo individuo puede aparecer registrado de formas diferentes. Tal es el caso de personas con doble nombre o apellido, que pueden aparecer alternativamente con cualquiera de ellos o con ambos al mismo tiempo.

- El problema complementario es el de la homonimia, cuando dos individuos diferentes tienen nombres exactamente iguales (o identificables como el mismo nombre).

Estos dos problemas hacen que la identificación no pueda efectuarse simplemente por coincidencia directa de nombres, y se deba recurrir a criterios más complejos.

Como primer paso en el proceso de vinculación, debe establecerse con claridad qué registros se desea vincular, y con qué objetivo. Si un objetivo es reconstruir la vida de un individuo y contamos con actas de bautismo, matrimonio y defunción, seguramente buscaremos vincular los registros identificando al individuo principal de cada uno de ellos. Si se trata de construir genealogías, puede interesarnos vincular al bautizado como padre en un matrimonio para descubrir a sus hijos, detectar su situación fiscal para relacionar datos como nivel económico y fertilidad, etc.

3.1 La construcción de genealogías en Sacaca y Acasio

En el caso de Sacaca y Acasio hemos trabajado únicamente con actas matrimoniales. El objetivo fue mencionado en la introducción: reconstruir genealogías

parciales, lo más extendidas posibles en el tiempo, que revelen la mayor cantidad de vínculos posibles —de afinidad o de consanguinidad— entre los contrayentes.

Cada acta identifica a los contrayentes, sus padres y sus parejas anteriores en el caso de los viudos. Más adelante veremos la composición exacta y completa de un acta, pero por el momento cabe identificar los casos de vinculación que interesan y los objetivos que cumple cada caso dentro del principal objetivo de la construcción de las genealogías de los contrayentes:

Vinculación Contrayentes - Padres de Contrayente:
Al identificar a una pareja de contrayentes en un acta A_1 como padres de otra pareja en otro acta A_2, se extiende la profundidad de la genealogía identificando 3 generaciones (los padres de los contrayentes en A_1, los contrayentes en $A_1 \equiv$ padres en A_2, y los contrayentes en A_2, nietos de los primeros).

Vinculación Padres de Contrayente - Padres de Contrayente:
Se intenta identificar a una pareja de padres de alguno de los contrayentes en un acta A_1, como padres en otro acta A_2. Esto permite extender a lo ancho la genealogía, identificando a dos de los contrayentes de las actas A_1 y A_2 como hermanos/as, o como el mismo individuo contrayendo matrimonio nuevamente.

Vinculación Contrayente - Contrayente:
La reaparición de un contrayente casándose nuevamente no extiende la genealogía, pero la identificación interesa a efectos de reconstruir la vida del individuo en cuestión (sus sucesivos matrimonios).

3.1.1 Composición del acta matrimonial

Describiremos la composición completa de un acta matrimonial de Sacaca y Acasio, para luego analizar los posibles criterios a utilizar en la vinculación.

El acta está compuesta por los siguientes datos (destacando que no siempre aparecen consignados todos ellos):

Día, mes y año de celebración del matrimonio.
Nombre y apellido del novio
Estado civil (Soltero o Viudo) del novio
Nombre y apellido de la novia
Estado civil (Soltera o Viuda) de la novia
Nombre y apellido de los padres del novio
Nombre y apellido de los padres de la novia
Nombre y apellido del cónyuge anterior del novio o de la novia, en el caso de los viudos
Ayllu del novio
Ayllu de la novia
Origen del novio y de la novia
Residencia del novio y de la novia
Categoría fiscal o étnica del novio y de la novia
Parroquia donde se celebra el matrimonio

Veremos en qué consisten algunos de los atributos anexos, lo que nos permitirá comprender, una vez más, por qué es necesario el conocimiento del historiador a lo largo de todas las etapas de la reconstrucción:

Ayllu - Obtención, pérdida y recuperación: El ayllu es un sub- agrupamiento de los indígenas de la zona, que corresponde de alguna manera a un sub-grupo étnico. El ayllu es un atributo que los individuos heredan de su padre, lo que permite inferir que el ayllu del novio que aparece en el acta es a su vez el ayllu del padre del novio. Por otro lado, puede ocurrir que la mujer, al casarse con un hombre de un ayllu diferente, sea absorbida por este nuevo ayllu. Los individuos que emigran, a su vez, pierden el ayllu convirtiéndose en forasteros, o eventualmente pueden ser absorbidos por uno de los ayllus de su nueva locación. De esta forma, al intentar identificar individuos, habrá que tener cierto cuidado al efectuar comparaciones de ayllu. Como se mencionó antes, además, los ayllus están agrupados en dos *mitades*.

Origen: Refiere a la región, estancia o pueblo del cual es originario el individuo. Nuestra base identifica casi 2000 valores diferentes en este campo.

Residencia: Refiere a la región, estancia o pueblo en la que reside el individuo en el momento del matrimonio.

Categoría: La situación fiscal de cada individuo depende de su origen étnico, su situación legal u otros atributos personales. Algunos valores posibles para la categoría son: *español, mulato, esclavo, forastero, agregado, mestizo, pardo libre, etc., etc.*

Parroquia: Es el nombre de la parroquia donde se celebra el matrimonio.

3.1.2 Valoración de los campos del acta matrimonial

Teniendo en cuenta la composición del acta, mencionaremos el grado de importancia otorgado a ciertos atributos y algunas observaciones que tuvimos en cuenta a la hora de definir los criterios de vinculación a utilizar.

El ayllu

Como mencionamos antes, existe la posibilidad de que un individuo pierda o cambie su ayllu a lo largo de su vida. La pérdida de ayllu parece ser una posibilidad más o menos frecuente, lo que lleva a que un individuo registrado como perteneciente a uno en un momento aparezca sin ayllu en un momento posterior. El cambio de ayllu, sin embargo, parece ser una posibilidad menos frecuente.

Por lo tanto, hemos adoptado dos posibles situaciones que consideraremos —con distinta valoración—, para posibles identificaciones: la coincidencia de ayllu, o la no contradicción. Consideraremos no contradicción el caso donde un individuo está identificado en un ayllu en un acta y no tiene ayllu en otra. Se tomó la decisión, entonces, de descartar identificaciones si el ayllu es contradictorio (descartando identificar posibles casos de cambio de ayllu, por ser infrecuentes y de difícil rastreo).

Origen y residencia

La gran variedad de lugares de origen y residencia posibles hacen que estos campos sean de poca confiabilidad a la hora de identificar individuos.

El lugar de residencia, en particular, puede variar numerosas veces a lo largo de la vida de un individuo. El lugar de origen es, en principio, más estable. Aún así, las variables formas de denominación de un mismo lugar y los distintos niveles de agregación con que se lo puede nombrar lo hacen de difícil valoración.

Para tomar en cuenta efectivamente estos campos, sería necesario establecer una jerarquía de inclusión entre las zonas geográficas, como se describió en el capítulo anterior.

En esta primer etapa, hemos decidido no considerar en absoluto el lugar de residencia como dato para la identificación. En el caso del origen, si coincide se considerará como un refuerzo del vínculo. Si no coincide, no afectará la decisión.

Veremos que en nuestro trabajo actual estamos incorporando la idea de considerar lugares de residencia particulares, muy reducidos, como criterio inicial importante para la identificación. Básicamente, se considera que durante un determinado período de tiempo, en un mismo lugar reducido, sólo habitaron un grupo pequeño de familias, y por lo tanto cualquier coincidencia nominal, en ese reducido espacio y tiempo, tiene altísima probabilidad de referir a un mismo individuo o a un integrante del mismo grupo familiar.

Categoría

La categoría podría variar, en principio, en los casos en que se refiere a una situación particular de un individuo (forastero, agregado). Cuando refiere a su origen étnico, el valor del dato es mayor (un español debería seguir siéndolo siempre). Sin embargo, sólo un 30% aproximadamente de las actas de matrimonio incluyen la categoría del novio, y un número similar se observa en el caso de las novias.

Hemos decidido tomar la categoría sólo como dato de refuerzo, principalmente para los españoles y mestizos, pero no como dato de identificación.

Las mujeres y la transmisión de apellidos

Un caso particular es el del nombre de las mujeres. Parece bastante claro que los habitantes de Sacaca y Acasio, o en todo caso los párrocos registrantes, se preocupaban bastante menos por la rigurosidad en el registro de los nombres femeninos que los masculinos. Por otro lado, se produce una situación particular en torno a la transmisión de los apellidos de padres a hijos.

Se ha mostrado que la aparición del *apellido*, a la usanza europea, fue consecuencia en los Andes de la conquista española (ver [21]). Fue común que, en el momento del bautismo, la hasta entonces única denominación indígena del individuo pasara a convertirse en su *apellido*, mientras se le agregaba un nombre de pila cristiano. Sin embargo, la práctica habitual europea de transmisión del apellido de padre a hijo no se impuso hasta avanzado el período de nuestro estudio.

Los registros sugieren que coexistieron, durante un tiempo prolongado, prácticas variadas de transmisión de apellidos, entre ellas (ver [1]):

- Transmisión estilo europeo del apellido paterno.
- Transmisión del apellido materno a las hijas, y paterno a los hijos.
- Utilización de un *apellido* distinto al de ambos padres. Se han detectado casos en los que varios hermanos comparten este *nuevo* apellido.

- Nombres de pila del padre que se convierten en apellido de los hijos.

En particular, sorprende la cantidad de mujeres que carecen de un apellido que las identifique, y sólo se las designa mediante un doble nombre de pila cristiano, como *María Rosa*, *Joana Isavel*, etc.

Estos mecanismos múltiples de transmisión de apellidos varían según el sexo: mientras que en el caso de los varones el 71% lleva el apellido de su padre, para las mujeres dicho porcentaje es sólo del 30% (ver [1]). A su vez un 5% de las mujeres lleva el apellido de su madre y, en el resto de los casos, o bien no se cita el apellido o bien el mismo no proviene de los padres.

Para empeorar la situación, en el caso de las mujeres que están nombradas sólo por su doble nombre de pila suele ocurrir que uno de ellos es `Maria`: el 56% del total de mujeres de nuestra base lleva éste como uno o el único de sus nombres.

Para subsanar en parte este problema, se definió para las novias el campo *Nombre completo*, que está constituído por el nombre y apellido[1] registrado de la novia, al que se le suma el apellido del padre y el apellido de la madre. El criterio de identificación nominal entre dos mujeres, entonces, se definió de la siguiente forma:

> Dos nombres de mujeres serán considerados equivalentes a efectos de su vinculación si coincide uno de los términos del nombre original que no sea `Maria`, más algún otro término de su *Nombre Completo*.

Un criterio similar, pero incorporando solamente el apellido del padre al nombre registrado, se aplicará a los novios en algunos casos de vinculación.

3.1.3 Criterios de vinculación

El proceso de vinculación de registros debe realizarse identificando reapariciones nominales de un individuo, y contando con al menos 2 ó 3 elementos extra de identificación suficientemente sólidos. Estos elementos pueden ser el nombre de los padres del individuo y/o el nombre de su pareja.

El problema de la homonimia existe, pero es claro que la probabilidad de encontrar dos parejas donde haya homonimia entre sus dos integrantes es muy baja. Sin embargo, se debe tener cuidado porque situaciones de este estilo pueden producirse, y sobre todo es factible confundir entre padres e hijos cuando llevan el mismo nombre. No habrá de sorprendernos encontrar una línea de abuelo-padre-nieto donde todos llevan el mismo apellido y al menos uno de sus nombres en común.

Tempranamente en el proceso de vinculación de registros, entonces, se definió que la identificación de individuos sueltos carece de sentido en este proyecto, y que cualquier identificación debería realizarse buscando reapariciones de parejas o de individuos con sus padres.

Teniendo en cuenta esto, existen 13 posibles vínculos a establecer:

[1] En los casos en que un individuo está registrado con dos nombres aparentemente de pila, el primero de ellos se ha cargado en el campo Nombre y el segundo en el campo Apellido.

Vincular...	*con...*
Pareja de contrayentes	*Padres de novio*
	Padres de novia
	Novio + Novia anterior
	Novio anterior + Novia
Padres de novio	*Padres de novio*
	Padres de novia
	Novio + Novia anterior
	Novio anterior + Novia
Padres de novia	*Padres de novia*
	Novio + Novia anterior
	Novio anterior + Novia
Novio + Padres	*Novio + Padres*
Novia + Padres	*Novia + Padres*

Vinculación Contrayentes - Padres de Contrayentes

Se trata de identificar a una pareja de novio y novia reapareciendo en un acta posterior como padres de otros contrayentes. El criterio a utilizar para identificar a los contrayentes como padres de novio o padres de novia es exactamente el mismo.

El primer criterio, para establecer un posible vínculo, es identificar nominalmente a cada novio con todas las reapariciones del mismo nombre como Padre de Novio (o Novia, según el caso). Utilizamos el nombre masculino como vinculador inicial teniendo en cuenta las dificultades antes mencionadas para los nombres femeninos.

Una vez identificada una posible relación, se la valida comparando el nombre de la novia con el de la madre de novio/a. El criterio a utilizar en esta validación es el mencionado al final de la sección anterior: debe coincidir algún elemento del nombre de la novia —distinto de María— con el de la madre del contrayente, más una coincidencia entre el *nombre completo* de la novia y el de la madre.

Al mismo tiempo se chequea que no exista contradicción de ayllu entre el novio y el padre de novio/a. Aunque el ayllu de los padres no figura en las actas matrimoniales, teniendo en cuenta que la transmisión de ayllu más fuerte se produce de padre a hijo/a, es posible considerar como ayllu del padre aquél que figura como el de su hijo/a. Por esta misma razón no es posible validar el ayllu de la novia contra el de la madre, porque no puede suponerse firmemente nada sobre el mismo en la segunda acta.

Por último, se valida que el matrimonio donde los contrayentes oficiarían de padres ocurra después del matrimonio propio, con una distancia de al menos 15 años. Suponiendo que los contrayentes, luego de casarse, tengan un hijo o hija, el mismo debería tener al menos unos 15 años de edad al casarse. De la misma forma, se valida que el matrimonio no ocurra más de 70 años después.

La figura 3.1 muestra un ejemplo de dos actas que resultaron vinculadas a través de este criterio. Podemos ver que el novio de la primer acta, *Andres Ayanuma*, reaparece como padre de novio en la segunda. A su vez, la novia *Asencia Maria*, hija de *Blas Mamani*, aparece registrada en el acta de la derecha como *Aciencia Mamani*. Es decir, además de la unificación ortográfica *Asencia-Aciencia*, se aplica el criterio del nombre completo para identificarla. Se puede

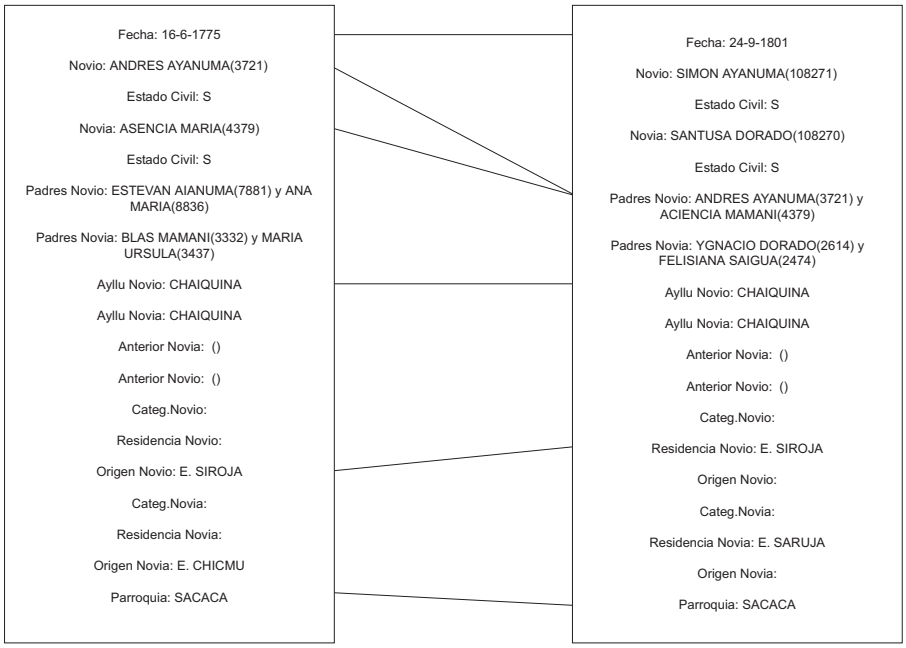

Figura 3.1: Dos actas con un vínculo entre los contrayentes y los padres del novio.

ver que los ayllus coinciden, lo que constituye un elemento más de identificación. Por otro lado, la distancia de 26 años entre el matrimonio de los contrayentes y el de sus hijos es perfectamente razonable. Por último, se puede ver que el origen del novio en el primer acta —*Estancia Siroja*— coincide con el lugar de residencia del novio en la segunda. Este elemento puede ser utilizado como reafirmante de la identificación en un eventual caso en que hubiera que decidir entre vínculos inconsistentes.

Vinculación Contrayentes - Matrimonios anteriores

En este caso, la vinculación se establece entre un matrimonio y el matrimonio anterior de alguno de los contrayentes. Se intenta vincular la pareja Novio-Novia con la pareja Novio-Anterior Novia o Novia-Anterior Novio.

En este caso también, el vínculo inicial se realiza entre los nombres masculinos (Novio vs Novio o Novio vs Anterior Novio). Luego, una primer validación se efectúa comparando los nombres femeninos bajo el criterio de comparación ya mencionado.

Teniendo en cuenta que uno de los contrayentes se está casando nuevamente, un importante elemento para su identificación podría ser el nombre de sus padres. Sin embargo, es muy común que para los viudos el nombre de los padres no se mencione. Aunque existen excepciones donde sí se mencionan, lamentablemente esta casi generalizada omisión del nombre del padre y la madre elimina una muy importante posibilidad de identificación.

El siguiente criterio será, nuevamente, el ayllu del novio o novia comparado

contra sí mismo en su nuevo casamiento.

Por último, se debe verificar que el segundo casamiento ocurra temporalmente después del primero. Aunque no hay límite inferior para la cantidad de años (alguien puede casarse en dos años seguidos), sí se establece un límite superior de unos 50 años.

Vinculación Padres de Contrayentes - Padres de Contrayentes

Como hemos mencionado, aquí se trata de vincular parejas de padres de novios/as entre sí, con el objetivo de identificar hermanos y extender horizontalmente las genealogías construídas.

Como siempre, el primer elemento para la vinculación será el nombre masculino. En este caso, el nombre del padre del novio/a con el nombre del padre del otro novio/a.

En segundo lugar, se valida el nombre de las madres. Cabe aclarar que, a diferencia de los nombres de las novias, el nombre de las madres suele ser un poco más estable e incluir su apellido. Es por ello que, en este caso, validamos el nombre de la madre considerándo solamente la coincidencia de un elemento distinto de María.

El último elemento identificatorio será también el ayllu de los novios/as, que al ser heredado de sus padres debería mantenerse.

Por último, en este caso también, aunque la distancia en años es potencialmente mucho más amplia, se establece un límite de unos 80 años.

Vinculación Padres de Contrayentes - Matrimonios anteriores

En este caso se busca vincular a una pareja de padres de contrayentes con un par novio/a-anterior novia/o.

Como en todos los casos anteriores, el vínculo inicial se establece entre los hombres. En este caso, entre el novio o anterior novio y el padre.

En segundo lugar, se valida el nombre de las mujeres con el criterio aplicado para las novias (utilizando el *nombre completo*) y se verifica la no contradicción de ayllu en el caso que corresponda (el ayllu del anterior novio/a no lo podemos determinar).

Aquí también es difícil establecer una ventana temporal, pero la hemos fijado en unos 80 años.

Vinculación Novios - Novios

Este último caso es particular: aquí intentamos identificar a un novio o novia suelto con alguna de sus reapariciones. Aunque esta identificación per se incluye a otras, como la de *Contrayentes - Matrimonios anteriores*, puede darse el caso —y de hecho hemos encontrado varios— donde un individuo se casa más de una vez pero no se encuentran todos los registros correspondientes y necesarios.

Puede darse el caso, además, de matrimonios sucesivos de individuos donde figuran siempre como solteros, posiblemente porque simplemente le mentían al párroco respecto a su estado civil.

Un ejemplo claro puede verse en la figura 3.2, donde *Martin Aguilario*, hijo en ambos casos de *Thomas Aguilario y Maria Barbara*, del ayllu *Sullcaticani*, originario de la *Estancia Charcoma*, se casa sucesivamente en los años 1777 y 1783 declarándose soltero en ambos casos.

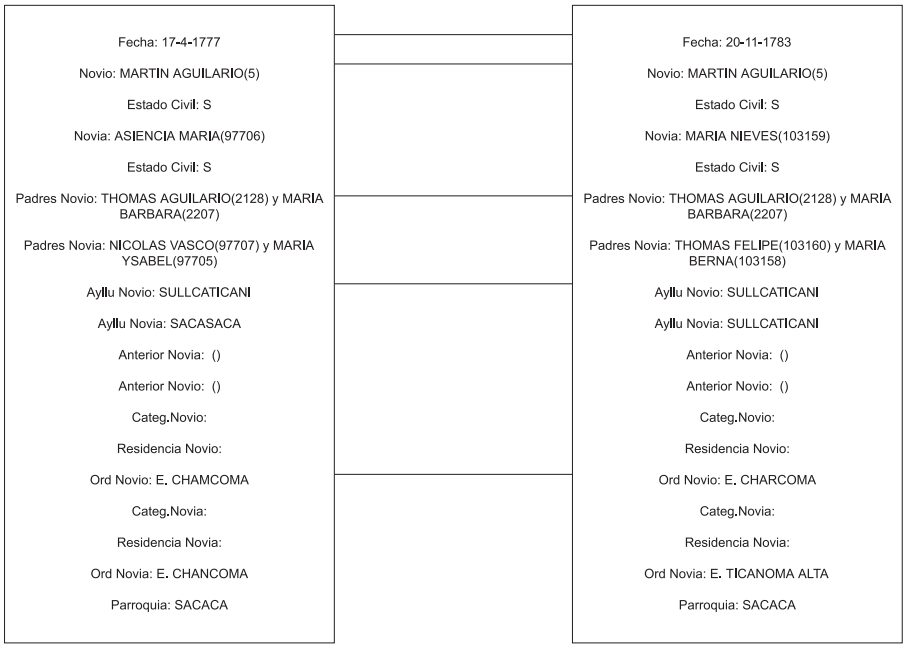

Figura 3.2: Dos actas con un vínculo entre los novios, ambos solteros!

La identificación, en el caso de los novios, comienza vinculando sus nombres en forma directa (tratándose de nombres masculinos). En el caso de las novias, se aplica el criterio de coincidencia parcial del *nombre completo* ya mencionado.

El principal elemento verificador es, luego, el nombre de sus padres. El problema aquí es que muchas veces, en el caso de los viudos, el mismo no figura.

Otro elemento verificador será obviamente el ayllu, que en este caso cobra bastante importancia. En segundo lugar, debe verificarse lo que llamamos coherencia de estado civil: nadie puede ser viudo primero y soltero después. Lo esperable serían casos de solteros que luego son viudos, o viudos que lo siguen siendo en sucesivos matrimonios. Hemos encontrado sin embargo casos, como el mencionado antes, de solteros que siguen siendo solteros.

Finalmente, establecimos una ventana temporal de 50 años entre sucesivos matrimonios.

3.2 Conclusiones generales

A partir del caso de Sacaca y Acasio, intentaremos plantear algunas observaciones y criterios generales para los procesos de identificación de individuos o vinculación de registros.

3.2.1 Criterios de vinculación múltiples

Los criterios antes mencionados son aquellos que finalmente tomamos como razonablemente *fuertes* para la identificación de individuos en Sacaca y Acasio.

A lo largo del proceso completo, sin embargo, muchas veces debe experimentarse con diferentes criterios. Si se establecen reglas demasiado rígidas, las identificaciones serán más *seguras*, pero se perderán muchos casos. Si las reglas se tornan demasiado laxas, la cantidad de vinculaciones crecerá pero generaremos muchos casos erróneos.

El balance entre optimismo y mesura suele ser delicado de encontrar. El exceso de mesura conduce a pobres resultados, y el exceso de optimismo puede generar una situación de demasiado ruido, con demasiadas inconsistencias.

Es conveniente ensayar con 2 o 3 criterios de vinculación, revisar manualmente los resultado obtenidos para detectar posibles errores, e ir iterativamente refinando el criterio hasta llegar a un punto aceptable.

Es en este proceso de revisión manual y refinamiento donde datos anexos —y generalmente difusos— toman mayor importancia. Frente a un caso de identificación dudosa o débil, una coincidencia en el origen de un individuo, o en su categoría fiscal, puede servir de refuerzo.

La herramienta desarrollada permite definir y revisar manualmente varios criterios diferentes para un mismo tipo de vinculación, marcando y comentando casos dudosos, o aceptables, etc.

3.2.2 Reglas generales de identificación

A partir de la experiencia de Sacaca y Acasio, hemos definido ciertos tipos de operadores y funciones booleanas que consideramos necesarios para cualquier conjunto de reglas de identificación.

Una regla de identificación es, en sí, una función booleana que recibe como parámetros dos conjuntos de atributos (dos actas matrimoniales, por ejemplo), y dos elementos particulares de dichos conjuntos (dos individuos que pueden o no ser el mismo), y devuelve verdadero cuando los individuos se identifican como el mismo.

La función en sí será una conjunción o disjunción de otras funciones booleanas, las cuales intentamos generalizar aquí:

Coincidencia Nominal Directa Dos nombres (nombre de pila simple o doble, más apellido) coinciden totalmente y en el orden en que se presentan.

Coincidencia Nominal Sin Orden Dos nombres coinciden completamente, sin importar el orden: `Joana Isavel` es equivalente a `Isavel Joana`

Coincidencia Nominal Parcial Coinciden algunos de los elementos de un nombre. Se debe especificar cuántos elementos se desea hacer coincidir: `Juan Diego Rivera de Solis` coincide en 3 elementos con `Diego Rivera Solis`.

Coincidencia Nominal Extendida Es la función utilizada en el caso de las mujeres. Recibe el nombre originalmente registrado de un individuo, el nombre extendido de dicha individuo (por ejemplo, incorporando apellidos de sus padres) y el nombre del individuo contra quien se lo compara, además de un conjunto de términos a excluir (por ej. `MARIA`). Devuelve verdadero si existe alguna coincidencia nominal entre los nombres originales, excluyendo los términos correspondientes, más otra coincidencia nominal utilizando el nombre extendido.

No Contradicción Compara dos términos verificando que, o bien coincidan, o bien uno sea nulo. Es el criterio utilizado para los ayllus.

No Contradicción Extendida Se puede extender el operador de No Contradicción para que, en el caso de que los dos términos sean no nulos, en lugar de coincidencia total se aplique alguna otra función booleana, como las definidas anteriormente.

Ventana Temporal Recibe dos años y un intervalo de tiempo. Devuelve si la diferencia entre el primero de los años y el segundo está dentro del intervalo en cuestión. Cabe remarcar que si lo que interesa es una distancia determinada, sin importar cuál de los eventos ocurrió primero, el intervalo debería comenzar con un valor menor a cero.

Par Válido Dado un par de elementos, y un conjunto de pares, devuelve verdadero si el par recibido pertenece al conjunto. Es el operador utilizado para la coincidencia de estados civiles. Si el año de un acta A_1 es menor que el año de un acta A_2, los pares ($<$*estado civil de*$A_1>$, $<$*estado civil de* $A_2>$) aceptados son $(Soltero, Viudo)$, $(Viudo, Viudo)$, $(Soltero, Soltero)$.

Inclusión Se utiliza para los casos denominaciones con distintos grados de agregación, como en las ya mencionadas referencias geográficas o de ocupación. A partir de una definición adecuada de los conjuntos, esta función recibe dos elementos y determina si son iguales, o uno puede incluir al otro. En el caso de Sacaca y Acasio, el operador se utiliza para comparar ayllus, donde en algunos casos la referencia en una acta puede hablar de una *mitad* que incluya al ayllu referido en el otro acta.

3.2.3 Operadores difusos, resultados difusos

De la definición anterior de los operadores, surge una primera posibilidad: convertirlos en operadores difusos. Es claro que no toda coincidencia nominal tiene el mismo peso: no es lo mismo una coincidencia nominal directa total que una coincidencia parcial, ni es lo mismo una coincidencia directa total entre los nombres `Juan Mamani` y `Diego Rodrigo Ruiz de Vaca`.

En primer lugar, se podría asignar a cada operador un intervalo de resultados posibles en $[0, 1]$. La forma de calcular el resultado puede depender de diversos factores:

- El valor devuelto por una Coincidencia Parcial podría depender de cuántos términos coinciden.
- Por otro lado, el valor de una coincidencia podría estar a su vez *pesado* por la frecuencia relativa de los términos que coinciden. Así, la coincidencia del término `Maria` no vale lo mismo que la coincidencia de `Villalobos`.
- Los años dentro de las ventanas temporales podrían tener una distribución particular (por ej., normal) que determine un *peso* diferente según la probabilidad de una diferencia dada de años entre registros.
- El operador Par Válido podrá tener distintos valores asociados a cada par. En nuestro ejemplo de los estados civiles, el par $(Soltero, Soltero)$ podría *pesar* menos que los otros.

Por otro lado, una función no necesariamente booleana podría determinar el *peso* (o probabilidad) total para la identificación que surge de una vinculación de dos actas. Podría determinarse que, por ejemplo, un valor 0 en una determinada función anula directamente la identificación (por ejemplo, una contradicción de ayllu), mientras que ciertos valores podrían sumar o restar al total con determinada ponderación (por ejemplo, una coincidencia en el nombre de un padre sumará más que la mera coherencia de estados civiles).

Del orden al caos: primer intento fallido en Sacaca y Acasio

Las posibilidades antes mencionadas, de definición de operadores difusos y de una función general de asignación de pesos o probabilidades a una identificación, fueron ensayadas sobre el caso de Sacaca y Acasio en un primer intento.

El primer obstáculo encontrado fue la dificultad de definir valores de retorno adecuados para cada función. Es muy difícil, a priori, responder preguntas tales como: ¿cuánto *vale* una coincidencia nominal en comparación con una coincidencia de origen? ¿Cuántas coincidencias en datos anexos compensan una contradicción en un apellido?

El segundo problema, suponiendo que se definan valores adecuados y razonables, es determinar los umbrales a partir de los cuales el valor asignado a un vínculo define una acpetación del mismo. Nuestra primer definición establecía un umbral debajo del cual los vínculos eran rechazados, uno a partir del cual se los aceptaba, y en el medio una llamada *zona gris*, para ser validada en forma manual.

La tercer dificultad está en definir una normalización adecuada que permita que los diferentes tipos de vinculaciones (contayentes-padres, padres-padres, etc.), donde la cantidad y tipo de operadores utilizados varía, puedan ser observados con el mismo criterio y los mismos umbrales. O, en todo caso, la dificultad se multiplica por la necesidad de definir umbrales diferentes según el tipo de vínculo.

La experiencia trabajando con este criterio nos llevó a ver que, a la larga, nos encontrábamos artificialmente asignando valores y umbrales de manera tal que los valores de aceptación respondieran a los criterios descriptos en la sección 3.1.3.

En definitiva, la conclusión a la que arribamos es casi una conclusión del terreno de la Ingeniería de Software: para el usuario es más sencillo y claro pensar en términos de criterios firmes de identificación, que pensar en una función de agregación de sub-criterios difusos. La alternativa propuesta en esta sección, aunque teóricamente correcta, demuestra ser poco práctica y utilizable. Algunos meses de trabajo y una serie de resultados bastante caóticos y ruidosos sostienen la base de esta conclusión, que hemos podido contrastar informalmente con algunos investigadores que han trabajado en procesos de reconstitución similares.

Cabe aclarar que esto no implica que, quizás, algunas ideas no sean utilizables. Aunque no lo hemos hecho en el presente trabajo, creemos que la posibilidad de incorporar a nuestras funciones booleanas algún elemento probabilístico (función de la frecuencia de algunos términos, por ejemplo) no debe ser descartada.

3.3 Inconsistencias

Una vez finalizada el proceso de aplicación de los criterios de vinculación, se genera implícitamente una genealogía. Como ya hemos mencionado, la identificación de un novio como padre nos agrega información sobre abuelos y nietos, la identificación de padres determina hermanos, etc.

Pero el proceso de identificación de pares puede llevar a inconsistencias. Como establece Wrigley en [41], para nuestra relativa tranquilidad, "[..] Perfect accuracy is beyond attainment in historical record linkage. [..]".

Cierta laxitud en los criterios conduce necesariamente a este tipo de errores. Por ejemplo, hemos detectado casos donde identificamos como el mismo individuo a un padre y su hijo, que presentan cierto grado de homonimia. Algo así puede ocurrir, quizás, porque en alguna de las actas no se menciona por ejemplo a la madre, y los criterios de no-contradicción permiten de todas formas la identificación.

Se decidió por lo tanto establecer los chequeos de consistencia como una etapa en sí misma, posterior a la etapa de vinculación de registros. El objetivo es mantener el proceso de vinculación lo más simple posible, para luego detectar y resolver manualmente los posibles errores.

Identificamos los siguientes tres tipos de inconsistencias que pueden producirse en una genealogía construída simplemente por identificación de individuos mirando los registros de a pares:

1. La aparición de individuos con más de un padre o madre.
2. Individuos que juegan más de un rol en un mismo acta. Es el caso que se produce cuando, a través de una cadena de identificaciones, un padre y un hijo se marcan como la misma persona. El resultado puede ser, luego, que en un acta de matrimonio el padre y el hijo tengan el mismo identificador.
3. Individuos que se casan dos veces con la misma persona. Al identificar a un individuo que se casa repetidas veces, puede ocurrir que un encadenamiento de vínculos nos hayan llevado a identificar también a dos novia diferentes como la misma persona.

Los casos de inconsistencias son chequeados sobre la genealogía generada, y los detectados deben ser expuestos al historiador para su resolución manual.

Cabe aclarar que otro tipo de inconsistencias pueden aparecer cuando no miramos las identificaciones de a pares sino en su conjunto. Por ejemplo, imaginemos un proceso de reconstrucción utilizando actas de bautismo, matrimonio y defunción. Podrá ocurrir que aceptemos que entre el bautismo y el matrimonio haya una distancia temporal de, por ejemplo, 40 años. Por otro lado, podemos aceptar una distancia entre matrimonio y defunción de, digamos, 60 años. Pero si el individuo en todos los casos es el mismo, se daría la situación donde la distancia entre bautismo y defunción es de 100 años, lo que lo hace más dudoso.

Si una inconsistencia fue generada por un encadenamiento de identificaciones que se consideran, de dos en dos, aceptables, es difícil establecer un criterio automático de decisión entre casos contradictorios y, por lo tanto, debe generarse un proceso de detección ad-hoc y un mecanismo de resolución manual.

Hemos desarrollado una herramienta para la revisión manual de inconsistencias, mostrando las diferentes identificaciones que condujeron a cada una de ellas y permitiendo al usuario tomar las decisiones necesarias.

Este proceso demostró ser costoso en términos de horas-hombre, y bastante engorroso al requerir numerosas inspecciones de situaciones que, en muchos casos, deben ser resueltas con cierto grado de arbitrariedad.

Capítulo 4

Métricas y resultados

La literatura existente sobre procesos de reconstrucción de familias suele indagar muy poco sobre la definición de métricas útiles para evaluar los resultados obtenidos. Cabe tener en cuenta que, dado que cada proyecto se basa en una fuente particular, con distintos grados de exactitud y completitud, y sobre poblaciones diferentes, es difícil establecer parámetros comparativos.

Hemos buscado establecer, por lo tanto, parámetros que nos permitieran medir el grado de avance de la reconstrucción, aún cuando no pudiéramos compararlo con otros proyectos o con variables poblacionales conocidas.

4.1 Dos métricas definidas

4.1.1 Tamaño de la población y cantidad de identificaciones

Es posible, y la literatura sí es abundante en ese sentido, estimar el tamaño de una población conociendo o estimando algunas variables tales como la tasa de natalidad y mortalidad, tasas de migración, esperanza de vida, etc. Enrique Tandeter y Mario Boleda han realizado tal estimación, utilizando el software *POPULATE* de McCaa y Pérez Brignoli, para la población de Sacaca y Acasio a lo largo del período en estudio.

En proyectos de reconstrucción de familias utilizando actas de bautismo, matrimonio y defunción, o utilizando censos, es posible por lo tanto comparar los resultados obtenidos, en términos del tamaño de la población, con las estimaciones realizadas por otro métodos. De esta forma, el grado de aproximación al parámetro poblacional *real* sería una medida útil para estimar el grado de éxito del proyecto.

En nuestro caso, hemos trabajado solamente con actas matrimoniales, construyendo genealogías pero sin reconstruir las estructuras familiares completas. Por ejemplo, no contamos con referencias a niños pequeños (que quizás nunca se casen, o mueran antes de llegar a la vida adulta), ni podemos estimar fechas de defunción de los individuos identificados. Por lo tanto, no contamos con datos para realizar comparaciones confiables con el tamaño de la población.

Sin embargo, en el momento de testear criterios de identificación, o de comparar el proceso finalmente presentado con el proceso inicial de reconstrucción

automática utilizando funciones de *peso*, ha resultado de suma utilidad medir la cantidad de identificaciones obtenidas.

La cantidad de individuos al empezar el proceso, sobre la base de las 11750 actas matrimoniales, y considerando toda identificación nominal como si fuera un individuo diferente, es de 66739 (lo que significa un promedio de $5,67$ individuos diferentes por acta). Una vez realizadas las identificaciones, esta cifra se reduce a 46265, lo que significa un número de 20474 identificaciones aceptadas ($66739 - 46265 = 20474$).

Utilizando el sistema de *pesos*, más allá del punto donde se establecieran los umbrales de aceptación / rechazo / zona gris, nunca se aceptaron más de 1000 o 2000 identificaciones.

4.1.2 Profundidad genealógica

Otra métrica de interés, para evaluar el grado de éxito de la reconstrucción, es la profundidad genealógica promedio obtenida.

Al comenzar el proceso sólo contamos con individuos sueltos, que no pertenecen a ninguna familia, no conocemos a sus padres, hijos o hermanos. Un individuo en estas condiciones tendrá profundidad genealógica 1.

Al final del proceso, a partir de las genealogías reconstruídas, podemos volver a medir la profundidad genealógica de cada individuo.

En genealogía existen tres formas diferentes de medir este parámetro:

- Por la línea *agnaticia*, se cuenta a partir del individuo (el *ego*) *subiendo* por sus ascendentes masculinos (si conocemos su padre, tendrá profundidad 2, si conocemos su abuelo paterno 3, etc.).
- Por la línea *uterina*, se cuenta a partir del *ego subiendo* por sus ascendentes femeninos.
- Por la línea *cognaticia*, se cuenta a partir del *ego subiendo* por cualquiera de sus ascendentes, contando la longitud del camino más largo que se conozca.

La figura 4.1 muestra un ejemplo de una pequeña genealogía y la forma de calcular la profundidad genealógica de un individuo con los tres criterios.

En el ejemplo, la profundidad genealógica del individuo 77 se calcula:

- A lo largo de la línea *agnaticia*, la profundidad es 3: 77 - 3881 - 6581.
- A lo largo de la línea *uterina*, es también 3: 77 - 4543 - 2483.
- A lo largo de la línea *cognaticia*, el camino más largo posible, la profundidad es de 4: 77, 4543 (madre), 2623 (padre), 10725 (madre) o 10016 (padre).

Utilizando estos criterios, entonces, el cuadro 4.1 muestra la cantidad de individuos, separados por sexo, con cada una de las profundidades genealógicas obtenidas (de 1 a 7). En un período de 120 años, es esperable no encontrar profundidades mayores a 7.

Se puede observar la aparición de numerosos individuos con profundidad hasta 3, lo que implica que conocemos a sus abuelos. Por otro lado, aunque

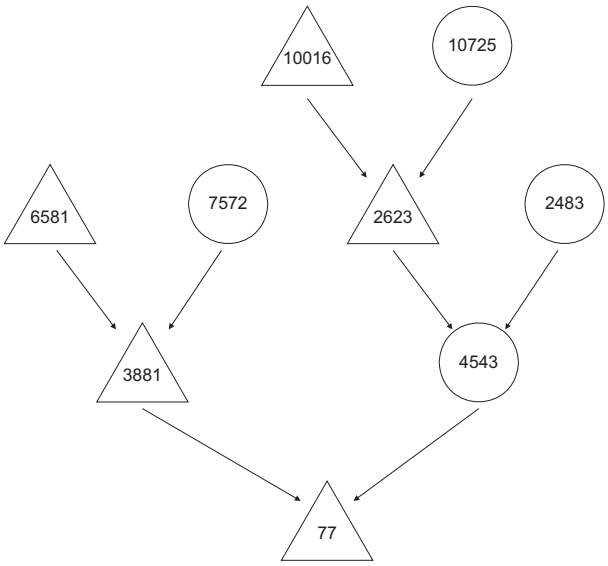

Figura 4.1: Ejemplo de una genealogía reconstruída. Cada individuo se identifica por un número único. Los triángulos representan hombres y los círculos mujeres.

Métrica	1	2	3	4	5	6	7
Hombres Agnaticia	2581 (22.38%)	7471 (64.77%)	1249 (10.83%)	195 (1.69%)	38 (0.33%)	1 (0.01%)	0 (0.00%)
Mujeres Agnaticia	2660 (23.49%)	7198 (63.55%)	1237 (10.92%)	198 (1.75%)	31 (0.27%)	2 (0.02%)	0 (0.00%)
Hombres Uterina	2497 (21.65%)	7423 (64.35%)	1365 (11.83%)	228 (1.98%)	22 (0.19%)	0 (0.00%)	0 (0.00%)
Mujeres Uterina	2557 (22.58%)	7296 (64.42%)	1243 (10.97%)	207 (1.83%)	21 (0.19%)	2 (0.02%)	0 (0.00%)
Hombres Cognaticia	2456 (21.29%)	7228 (62.66%)	1254 (10.87%)	430 (3.73%)	123 (1.07%)	36 (0.31%)	8 (0.07%)
Mujeres Cognaticia	2527 (22.31%)	7048 (62.23%)	1185 (10.46%)	387 (3.42%)	135 (1.19%)	36 (0.32%)	8 (0.07%)

Tabla 4.1: Número y porcentaje de individuos con profundidades genealógicas de 1 a 7.

es pequeña en términos absolutos, la cantidad de individuos con profundidad 4 (y hasta 5) resulta no despreciable en un proyecto de estas características. Debe tenerse en cuenta que un individuo con profundidad 4 sólo puede aparecer casándose unos 80 años después del inicio del período en estudio (que es de 120 años), lo que reduce significativamente el universo de individuos que potencialmente pueden tener esta profundidad genealógica.

4.2 Análisis de la genealogía resultante

Uno de los objetivos del proyecto de reconstrucción consiste en detectar estrategias matrimoniales de cierto grado de complejidad. Uno de estos análisis fue realizado utilizando el programa *GENOS*, desarrollado por Laurent Barry en el Laboratorio de Antropología Social del Collège de France en Paris. *GENOS* recibe como entrada una genealogía y analiza las estrategias subyacentes, en

términos de relaciones de parentesco por consanguinidad o afinidad, entre los contrayentes.

GENOS detecta dos tipos de estrategias: los llamados *redoblamientos* dobles y triples. Un *redoblamiento* doble consiste, básicamente, en una relación de consanguinidad o afinidad, en cierto grado, entre dos contrayentes. En un *redoblamiento* triple, tres grupos o familias intervienen. Un individuo de la familia A se casa con alguien de la familia B. A su vez, alguien de la familia B se casa con alguien de la familia C y, por último, alguien del grupo A se casa con alguien del grupo C.

Hemos desarrollado un visualizador propio que extiende la salida provista por *GENOS*, graficando los *redoblamientos* y mostrando en un mismo color a los individuos que pertenecen al mismo *ayllu*. El objetivo es analizar y visualizar rápidamente estrategias de intercambio inter e intra *ayllu*.

La figura 4.2 muestra el visualizador desarrollado, con un *redoblamiento* triple, donde se puede ver el esquema de alianzas e intercambios entre los 3 grupos involucrados, que en este caso además pertenecen todos al mismo *ayllu*.

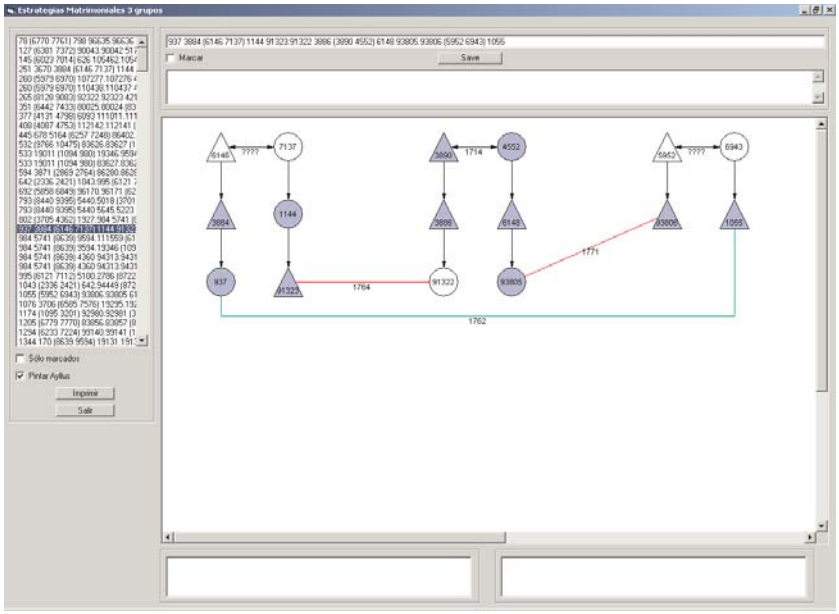

Figura 4.2: Visualizador de estrategias matrimoniales entre 3 grupos, a partir de la salida generada por GENOS.

Capítulo 5

Conclusiones, trabajo actual y futuro

5.1 Conclusiones

En primer lugar, creemos que la descomposición del proceso de reconstrucción en las etapas propuestas es de carácter general, aplicable a cualquier proyecto de este tipo, y que al identificarse y separarse de esta forma da lugar a resultados confiables y estables.

Por otro lado, creemos que el grado de generalidad y aplicabilidad de la metodología y las herramientas está dado, en gran medida, por el hecho de que permiten al historiador participar de las diferentes etapas incorporando su conocimiento contextual y su juicio, testeando y validando hipótesis.

Siguiendo a Adman (ver [2]), sostenemos que el rol del historiador es irremplazable a lo largo del proceso de reconstrucción, y que su juicio es imposible de transferir completamente a un proceso 100% automático. Por lo tanto las herramientas deben estar preparadas para ser configuradas adecuadamente en los puntos necesarios.

El conocimiento contextual fue clave, en el proceso de Sacaca y Acasio, a la hora de configurar los criterios de identificación y la homogeneización de los apellidos. La participación de historiadores y lingüístas determinó la configuración del sistema teniendo en cuenta los mecanismos de transmisión de apellidos, el proceso de herencia y pérdida de ayllus, el agrupamiento de ayllus en mitades, los factores lingüísticos intervinientes en los nombres, etc.

Por último, nos interesa marcar el carácter interdisciplinario del proceso de reconstrucción de Sacaca y Acasio, y resaltar la importancia que tuvo en su éxito la capacidad de cada disciplina de escuchar, comprender y adaptarse a las necesidades reales del proyecto. En palabras de John Jeacocke:

"Computing science is a young discipline and history is an old one; it is only recently that the special problems of the historian have begun to be looked at by computer scientists. [..] Impatient historians will only 're-invent the wheel'. Computer scientists who are unwilling to listen will invent tools that no one can use" (ver [16]).

5.2 Trabajo actual

En este momento, se está extendiendo el proyecto de Sacaca y Acasio mediante la incorporación de actas de bautismo y defunción.

Actualmente se está en proceso de digitalización de los registros, y de testeo preliminar de los criterios de identificación que serán aplicados.

Se está estudiando, a su vez, de qué manera mejorar y agregar funcionalidad y facilidad de uso a las herramientas existentes, aunque cabe destacar que hemos detectado que las herramientas actuales cubren las necesidades de este nuevo proceso.

5.3 Trabajo futuro

Creemos que una de las principales tareas hacia el futuro es la unificación de las herramientas en un paquete distribuible y realmente *amigable*.

Una de las principales debilidades actuales es la falta de una herramienta realmente flexible y sencilla de usar para la definición, configuración y testeo de los criterios de identificación. En la actualidad, la definición de estas reglas requiere de un grado de entrenamiento y/o la participación del desarrollador.

Un paquete de este tipo puede ser desarrollado y debería satisfacer las necesidades de cualquier familia de fuentes históricas que se desee vincular.

Por último, es necesario analizar la forma de incluir inteligentemente factores de probabilidad en los elementos de identificación. Si una categoría fiscal, un origen, e incluso un ayllu o un nombre son de baja frecuencia dentro de la base, su coincidencia es de mayor valor. Debe incorporarse este factor sin generar un ruido innecesario, tal como ocurriera con la utilización de la función de *peso*.

Bibliografía

[1] Luis Acosta and Enrique Tandeter. La transmisión de apellidos entre los indígenas andinos, siglos XVII-XIX. *Anuario 2002 del Archivo y Biblioteca Nacionales de Bolivia*, pages 355–369, 2002.

[2] Peter Adman, Stephen W. Baskerville, and Katharine F. Beedham. Computer-assisted record linkage: or how best to optimize links without generating errors. *History and Computing*, 4(1):2–15, 1992.

[3] Marc Bloch. Classification et choix des faits en histoire économique. *Annales d'Histoire Économique et Social*, 1929.

[4] André Burguière. *La historiografía francesa contemporánea*, chapter Historia de una historia: el nacimiento de Annales, pages 79–100. Editorial Biblos, Buenos Aires, 1990.

[5] Peter Burke. *La revolución historiográfica francesa. La Escuela de los Annales: 1929-1989*. Editorial Gedisa, Buenos Aires, 1993.

[6] Emile Durkheim. Cours de science sociale, leçon dóuverture. *Revue internationale de l'enseignement*, 15:23–48, 1888.

[7] Lucien Febvre. *Combates por la historia*. Ariel, Barcelona, 1953.

[8] François Furet. *L'atelier de l'Historire*. Flammarion, Paris, 1982.

[9] Jean Gaudemet. *Le mariage en Occident: les moeurs et le droit*. Les Editions du Cerf, Paris, 1987.

[10] Pierre Goubert. *Beauvais et le Beauvaisis de 1600 à 1730, contribution à l'histoire sociale de la France du XVIIe*. Editions de l'EHESS, Paris, 1960.

[11] Patrick Hanks and Flavia Hodges. *Dictionary of surnames*. Oxford University Press, Oxford, 1988.

[12] Louis Henry. *Manuel de démographie historique*. Flammarion, Paris, 1976.

[13] Françoise Héritier. "Parentela". *Enciclopedia Einaudi*, 10:394–399, 1980.

[14] Françoise Héritier. *L'exercice de la parentè*. Hautes Etudes/Gallimard/Le Seuil, Paris, 1981.

[15] Georg G. Iggers. *Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge*. Wesleyan University Press, Connecticut, USA, 1997.

[16] John Jeacocke. *Historians, Computers and Data*, chapter The Computer Scientist and the Historian, pages 39–44. Manchester University Press, Manchester, UK, 1990.

[17] D. E. Knuth. *The Art of Computer Programming. Volume 3: Sorting and Searching*. Addison-Wesley, Reading, MA, USA, 2nd edition, 1998.

[18] Ernest Labrousse. *Esquisse du mouvement des prix et des revenus en France au XVIII siècle*. Dalloz, Paris, 1933.

[19] Karl Lamprecht. *Deutsche Geschichte*. Berlin, 1891.

[20] Emmanuel Le Roy Ladurie. *Le territoire de l'historien, I*, chapter L'historien et l'ordinateur, pages 11–22. Gallimard, Paris, 1973.

[21] Ximena Medinacelli. ¿Nombres o apellidos? El sistema nominativo indígena en Sakaka en el siglo XVII. Master's thesis, Universidad Internacional de Andalucía, La Rábida, 1997.

[22] R.J. Morris. Editorial - nominal record linkage: into the 1990s. *History and Computing*, 4(1):iii–vii, 1992.

[23] Frankie Patman and Leonard Shaefer. Is Soundex good enough for you? On the hidden risks of Soundex-based name searching. Whitepaper, http://www.onomastix.com, 2001.

[24] Tristan Platt. *Estado boliviano y ayllu andino. Tierra y Tributo en el Norte de Potosí*. I.E.P., Lima, 1982.

[25] Jacques Revel. *Las construcciones francesas del pasado*. Fondo de Cultura Económica, Buenos Aires, 2001.

[26] Daisy Rípodas Ardanaz. *El matrimonio en Indias. Realidad social y regulación jurídica*. FECIC, Buenos Aires, 1977.

[27] Marion Selz-Laurière. Parente et informatique. *Mathématique et sciences humaines*, (97), 1987.

[28] Marion Selz-Laurière. Les mathématiques en ethnologie. *L'Homme*, XXVIII(4), 1988.

[29] Marion Selz-Laurière. Donnée de sciences humaines et intelligence artificielle. *L'Homme*, XXX(4), 1990.

[30] Marion Selz-Laurière. Informatique et sciences humaines: formalisation et démarche d'explicitation. *Gradhiva*, (14), 1993.

[31] Marion Selz-Laurière. Traitement informatique de données généalogiques: le logiciel 'gen-par'. *L'Homme*, XXXIV(2), 1994.

[32] Marion Selz-Laurière. Informatique, généalogies, parente. *Le Médéviste et l'ordinateur*, (36), 1997.

[33] Marion Selz-Laurière and Pierre Lamaison. Généalogies, alliances et informatique. *Terrain*, 1985.

[34] M. Skolnick, editor. *Conference on Methods of Automatic Family Reconstitution*, Liège, Belgium, 1978. International Union for the Scientific Study of Population.

[35] Alan Stanier. How accurate is SOUNDEX matching? *Computers in Genealogy*, 3(7), 1990.

[36] Enrique Tandeter. *Coacción y Mercado. La minería de la plata en el Potosí colonial, 1692-1826*. Editorial Sudamericana, Buenos Aires, 1992.

[37] Enrique Tandeter. Población y economía en los andes (siglo xviii). *Revista Andina*, 25(13):7–42, 1995.

[38] Merry E. Wiesner-Hansk. *Christianity and Sexuality in the Early Modern World. Regulating Desire, Reforming Practice*. Routledge, London.

[39] E. A. Wrigley, R. S. Davies, J. E. Oeppen, and R. S. Schofield. *English population from family reconstitution, 1580-1837*. Cambridge University Press, Cambridge, 1997.

[40] E. A. Wrigley and R. S. Schofield. *The population history of England 1541-1871: a reconstruction*. Cambridge University Press, Cambridge, 1989.

[41] E.A. Wrigley. *Identifying People in the Past*. Edward Arnold Publishers Ltd., London, 1973.

[42] R.T. Zuidema. *Andean Kinship and Marriage*, chapter The Inca Kinship System: A New Theoretical View, pages 240–281. American Anthropological Association, Washington, DC, 1973.

[43] R.T. Zuidema. *Le Nouveau Monde, mondes nouveaux : l'experience américaine*, chapter The Spanish Contributions to the study of Amerindian kinship system, pages 643–664. Editions Recherche sur les civilisations/Editions de l'Ecole des hautes etudes en sciences sociales, Paris, 1996.