IDB Inter-American Development Bank

Servicio Nacional de Migraciones
Ministerio del Interior
Gobierno de Chile

ESTADÍSTICAS MIGRATORIAS

Sistema Integrado de Información en Migración y Extranjería (SIIME)

Recomendaciones para el diseño del SIIME

2

Diciembre 2025

MIGRACIONES CHILE

CONTEXTO DEL PROYECTO

Este proyecto fue desarrollado gracias a la Cooperación Técnica (CH-T1287) del Banco Interamericano de Desarrollo (BID) con el Servicio Nacional de Migraciones (Sermig), iniciada en diciembre del año 2022.

El proyecto fue ejecutado por la consultora Belén Harnecker Fontecilla bajo la dirección del Departamento de Estudios del Sermig y con retroalimentación de distintas áreas del Instituto Nacional de Estadísticas (INE). El proyecto, se desarrolló bajo el alero de la agenda de trabajo de la Mesa de Estadísticas Migratorias, coordinada por el INE y el Sermig, período 2022-2025.

Agradecemos a todas y todos las y los profesionales del Estado de Chile que colaboraron en las distintas etapas de implementación del presente proyecto.

GLOSARIO DE SIGLAS

SERMIG: Servicio Nacional de Migraciones

SIIME: Sistema integrado de información en migración y extranjería

INE: Instituto Nacional de Estadísticas

ADR UK: Administrative Data Research UK

Sernameg: Servicio Nacional de la Mujer y Equidad de Género

MINEDUC: Ministerio de Educación

MINSAL: Ministerio de Salud

MDSF: Ministerio de Desarrollo Social y Familia

REP: Registro Estadístico de Población

RNE: Registro Nacional de Extranjeros

CONTENIDO

Introducción

En Chile, los registros administrativos constituyen una fuente fundamental de información para la gestión pública y la generación de estadísticas. Sin embargo, la ausencia de estándares comunes y de mecanismos de interoperabilidad limita su potencial, especialmente en el ámbito de la migración internacional, donde la coordinación entre instituciones es clave para garantizar la eficiencia institucional, la calidad de la información y la protección de derechos de las personas migrantes.

En este contexto, se propone el desarrollo de un Sistema Integrado de Información en Migración y Extranjería (SIIME), iniciativa conjunta del Servicio Nacional de Migraciones (SERMIG) y el Banco Interamericano de Desarrollo (BID). El SIIME responde a los diagnósticos identificados en el *Mapa de Registros* (primer producto del proyecto), que evidenció la fragmentación y heterogeneidad de los datos personales de personas extranjeras en los registros del Estado, así como la necesidad de contar con lineamientos comunes que orienten su recolección (producto 2), intercambio e integración (producto 3).

El SIIME tiene una triple finalidad:

a. Calidad estadística
b. Generación de evidencia para formulación y monitoreo de políticas públicas
c. Eficiencia institucional: validación con el RNE (diagnóstico producto 2)

El presente informe tiene por objetivo establecer los lineamientos metodológicos y técnicos para el diseño del SIIME. Se organiza en seis secciones: (i) relevancia de contar con un sistema integrado y potencial de uso en Chile; (ii) marco metodológico, que define principios, protocolos de seguridad y gobernanza; (iii) marco técnico, que describe alternativas de integración, incluyendo metodologías de vinculación de datos; (iv) consideraciones legales y de protección de datos; (v) recomendaciones y próximos pasos; y (vi) conclusiones.

1. Relevancia del SIIME

El diseño del SIIME responde no solo a un imperativo de eficiencia administrativa, sino también a la necesidad de fortalecer la calidad de las estadísticas nacionales, optimizar la gestión pública y garantizar los derechos fundamentales de las personas migrantes.

La actual fragmentación de los registros administrativos en Chile produce duplicidades, vacíos y dificultades para obtener información confiable, lo que limita la capacidad del Estado para diseñar políticas basadas en evidencia. Este desafío es compartido internacionalmente: organismos como la UNSD han señalado que la integración de registros es condición necesaria para avanzar hacia estadísticas migratorias oportunas y comparables (UNSD, 2025).

El SIIME permitirá superar estas limitaciones al integrar diversas fuentes bajo una gobernanza común. De este modo, el sistema aportará valor en tres dimensiones estratégicas: calidad estadística, generación de evidencia para políticas públicas y eficiencia institucional.

2.1 Calidad estadística

La integración de registros provenientes de distintas instituciones incrementará la calidad de las estadísticas existentes y abrirá la posibilidad de generar indicadores inéditos. Experiencias como la estimación conjunta de población extranjera por INE y SERMIG ya han mostrado que la combinación de mejora la cobertura y visibiliza tanto a la población regular como irregular.

Además, el SIIME permitirá construir nuevos indicadores. Un ejemplo de indicador interinstitucional ya implementado es la tasa de cobertura de permisos humanitarios por violencia intrafamiliar (VIF) elaborada por SERMIG y SERNAMEG, que evidencia brechas críticas en la regularización de mujeres extranjeras víctimas de violencia de género. Indicadores como éste muestran cómo la integración de registros puede contribuir a la identificación de poblaciones vulnerables y orientar políticas focalizadas. En la Tabla 1 se presentan ejemplos adicionales de indicadores que podrían desarrollarse mediante la cooperación institucional y el uso integrado de registros administrativos.

Por otro lado, de cara al futuro, tras el Censo 2024, será necesario establecer una metodología estable de estimación de población extranjera basada en registros administrativos. El SIIME puede convertirse en la infraestructura clave para esa transición.

Tabla 1: Ejemplos de bases de datos e indicadores que se podrían construir a partir de la vinculación de registros administrativos entre diferentes instituciones

Tema	Base de datos / objeto de análisis	Fuentes de datos requeridas	Valor agregado de la integración	Ejemplo de indicadores
Salud	Acceso a servicios de salud	Registros de salud (MINSAL) + Estatus migratorio (SERMIG)	Permite identificar brechas de acceso según estatus migratorio y tipo de residencia.	• Tasa de acceso a controles de embarazo en mujeres migrantes. • Tasa de cobertura de permisos humanitarios por embarazo • Porcentaje de migrantes afiliados a FONASA
Trabajo	Participación laboral formal e informal	Registros laborales (DT, AFC, SII) + Nacionalidad y estatus migratorio (SERMIG)	Facilita comparar condiciones laborales entre población nacional y migrante.	• Tasa de empleo formal de personas migrantes. • Brecha de ingresos medios entre migrantes y nacionales.
Educación	Trayectorias educativas de NNA migrantes	Registro de matrícula escolar (MINEDUC) + Estatus migratorio (SERMIG)	Posibilita medir abandono escolar, sobreedad y continuidad educativa.	• Tasas de promoción, repitencia y retiro en estudiantes migrantes según estatus migratorio. • Tasa neta de escolarización de NNA migrantes según estatus migratorio. • Tasas de asistencia según estatus migratorio.
	Resultados educacionales y trayectorias postsecundarias	Matrícula escolar (MINEDUC) + Matrícula de educación superior (MINEDUC) + Estatus migratorio (SERMIG) + Registros laborales (DT, SII)	Permite realizar estudios longitudinales sobre educación e inserción laboral.	• Porcentaje de estudiantes migrantes que acceden a educación superior. • Inserción laboral de egresados migrantes a 1 y 3 años de titulación.
Protección social	Acceso a programas sociales	Registro Social de Hogares (MDSF) + Estatus migratorio (SERMIG)	Permite evaluar brechas en prestaciones sociales según situación migratoria.	• Tasa de acceso a subsidios familiares en hogares migrantes.

				• Porcentaje de hogares migrantes incorporados al RSH.
Vivienda	Acceso a subsidios habitacionales	Registros MINVU + Estatus migratorio (SERMIG)	Evalúa acceso a beneficios habitacionales y su distribución en población migrante.	• Tasa de adjudicación de subsidios habitacionales en personas migrantes con residencia permanente. • Porcentaje de hogares migrantes en situación de hacinamiento.
Justicia y protección	Violencia de género y VIF en mujeres migrantes	Registros SERNAMEG + Estatus migratorio (SERMIG)	Apoya políticas de protección y prevención con enfoque diferenciado.	• Tasa de cobertura de permisos humanitarios por VIF. • Porcentaje de mujeres migrantes víctimas de VIF que acceden a programas de protección.
Movilidad interna	Migración interna dentro de Chile (cambios de residencia)	Registros de residencia (MDSF) + Datos migratorios (SERMIG)	Permite identificar dinámicas de movilidad territorial y necesidades de planificación local.	• Tasas de migración interna de población extranjera por región. • Identificar regiones y/o comunas receptoras y expulsoras de población extranjera en Chile.

2.2 Generación de evidencia para la formulación y monitoreo de políticas públicas

Más allá de la estadística descriptiva, el SIIME podría transformarse en una plataforma de evidencia para la toma de decisiones. Al permitir el seguimiento longitudinal de trayectorias individuales, ofrecerá la posibilidad de analizar cómo evolucionan las condiciones de inclusión social de la población migrante (por ejemplo, en educación, salud o empleo) durante sus primeros años de residencia.

La información integrada también mejorará la focalización de programas sociales, al identificar territorios o grupos con mayores niveles de vulnerabilidad, como por ejemplo hogares migrantes en situación de hacinamiento, lo que habilita intervenciones más eficaces en vivienda o subsidios de arriendo.

Asimismo, el SIIME reforzaría la capacidad del Estado para monitorear compromisos internacionales: permitiría, por ejemplo, reportar indicadores ODS con desagregaciones por nacionalidad o condición migratoria, y cumplir con las recomendaciones de la UNSD sobre estadísticas migratorias.

La integración de datos facilitará la detección temprana de grupos vulnerables, como niños, niñas y adolescentes migrantes en situación irregular. Por ejemplo, al cruzar información del SERMIG con el Servicio Nacional de Protección Especializada de la Niñez y Adolescencia, podría ser posible identificar riesgos y garantizar acceso a derechos básicos.

2.3 Eficiencia institucional

La fragmentación actual se traduce en procesos administrativos duplicados, validaciones manuales y mayores costos de gestión. Durante el levantamiento del Mapa de Registros Administrativos del proyecto SIIME, diversas instituciones señalaron la falta de herramientas para validar datos de personas sin RUN.

El SIIME abordará este problema mediante la validación automática frente al RNE, permitiendo a instituciones como el Ministerio de Educación verificar en línea la situación de estudiantes extranjeros y reducir errores en matrícula o asignación de recursos.

La coordinación interinstitucional también se verá fortalecida: por ejemplo, los cambios en el estatus migratorio de una persona (por ejemplo, de residencia temporal a definitiva) podrían ser comunicados automáticamente a múltiples servicios, generando eficiencia operativa y evitando cargas innecesarias para las familias.

Infraestructuras de Datos Integrados (IDI): el caso de ADR UK (Reino Unido)

ADR UK (Administrative Data Research UK) es una iniciativa creada en 2018 para facilitar la investigación con datos administrativos, con el fin de generar evidencia útil para el diseño de políticas públicas y la toma de decisiones.

Está integrada por cuatro asociaciones nacionales (Inglaterra, Escocia, Gales e Irlanda del Norte) y coordinada por la Oficina Nacional de Estadísticas (ONS) junto al Consejo de Investigación Económico y Social (ESRC). La ONS lidera la identificación y vinculación de registros públicos y garantiza el acceso seguro de investigadores acreditados a través del *Secure Research Service (SRS)*.

En su etapa piloto (2018–2022) recibió £59 millones del ESRC y, desde 2021, cuenta con financiamiento estable hasta 2026. Para 2023 acumula más de 400 bases de datos, 7.000 investigadores acreditados, 1.300 proyectos desarrollados, 14 ministerios participantes y fondos concursables por £19 millones.

Un pilar central es la seguridad y privacidad: el acceso se realiza únicamente en entornos de investigación confiables (*Trusted Research Environments – TRE*), bajo el marco de los *Five Safes*. Investigadores y proyectos deben superar un estricto proceso de acreditación para trabajar con datos no identificados.

Otro eje clave es la confianza ciudadana. ADR UK mantiene programas de participación pública (paneles, charlas, encuestas) que aseguran que el uso de datos responda al interés público. La evidencia muestra que la ciudadanía apoya estas iniciativas si cumplen tres condiciones: que los proyectos aporten al bien común, que los datos sean protegidos de manera segura y que se actúe con transparencia.

En conjunto, ADR UK representa un modelo consolidado de gobernanza, seguridad y confianza para aprovechar el valor de los datos administrativos en beneficio de la sociedad.

Recuadro 1: Ejemplo internacional en infraestructuras de datos integradas, Administrative Research Data UK - ADR UK

3. Marco metodológico

El diseño del SIIME requiere un marco metodológico que defina con claridad cómo se recopilarán, intercambiarán y usarán los datos administrativos, garantizando a la vez la protección de los derechos de las personas migrantes y el resguardo de la información sensible. La experiencia internacional muestra que los sistemas de integración de datos deben basarse en principios como el de proporcionalidad, legitimidad, seguridad y transparencia (descritos en informe del producto 2).

Un referente relevante es el modelo de ADR UK que opera bajo un esquema de "federación controlada" de datos: la información permanece bajo custodia de los organismos que la generan, pero puede ser vinculada y analizada a través de protocolos estandarizados, en un entorno seguro y gestionado por un ente coordinador. Esta lógica se puede adaptar al caso chileno, donde el SIIME debe respetar la autonomía institucional, pero ofrecer mecanismos confiables de interoperabilidad y uso estadístico de los datos.

3.1 Principios metodológicos

El diseño metodológico del SIIME debe apoyarse en principios que aseguren un equilibrio entre la utilidad administrativa y estadística de los registros y la protección de los derechos de las personas migrantes. La literatura internacional subraya que la integración de datos solo es sostenible cuando se implementa bajo marcos de interoperabilidad, seguridad y gobernanza claros, que fortalezcan la confianza institucional y social (UNECE, 2011; OECD, 2021; ADR UK, 2021).

Asimismo, experiencias como el *Integrated Data Infrastructure* de Nueva Zelanda y los sistemas federados del Reino Unido evidencian que la claridad en los principios metodológicos es clave para lograr legitimidad, transparencia y consistencia en el uso de registros administrativos (Stats NZ, 2017; DARE UK, 2023).

A continuación, se proponen cinco principios metodológicos que podrían regir al SIIME, inspirados en estos ejemplos internacionales.

a. **Interoperabilidad regulada.**

   Los datos permanecen bajo custodia de cada institución, pero se comparten siguiendo estándares y reglas comunes. Esto incluye catálogos de variables y metadatos armonizados, el uso de codificaciones internacionales (como ISO 3166-1 para países) y protocolos de interoperabilidad (ej. APIs o envíos periódicos de archivos en formatos estandarizados). De esta manera, se garantiza que las instituciones "hablen el mismo idioma" al intercambiar información y se evitan duplicidades o inconsistencias.

b. Confidencialidad y seguridad reforzada.

El acceso a los datos debe realizarse bajo marcos de seguridad reconocidos, como los Cinco Safes (proyectos, personas, datos, entornos y salidas seguras). Esto implica aplicar cifrado en tránsito y en reposo, autenticación multifactor para usuarios, y el uso de entornos seguros de investigación (como los *Trusted Research Environments – TRE, del caso Reino Unido*) donde se pueda analizar información sin exponer microdatos.
Adicionalmente, toda salida debe pasar por procesos de revisión, y los sistemas deben contar con evaluaciones periódicas de impacto en privacidad (PIA).

c. Gobernanza compartida.

La coordinación interinstitucional requiere definir roles claros. Un comité interinstitucional debe establecer prioridades y normas de acceso, una unidad técnica garantizará la calidad y estandarización de los procesos, y un tercero de confianza (ej. INE o SERMIG) será responsable de recibir datos identificados, ejecutar procesos de vinculación y entregar datasets anonimizados a los usuarios autorizados. Este modelo asegura equilibrio entre autonomía institucional, neutralidad técnica y protección de los datos.

d. Minimización de datos.

El intercambio de información debe limitarse estrictamente a las variables necesarias para cada finalidad. Esto se operacionaliza mediante la creación de capas de datos: a nivel administrativo (con identificadores), a nivel analítico (con seudónimos) y a nivel de difusión (con datos agregados). Asimismo, los datos temporales utilizados para análisis deben ser eliminados una vez finalizado el proyecto, evitando acumulación innecesaria de información sensible.

e. Trazabilidad y transparencia.

Todo acceso, consulta o modificación de los datos debe quedar registrado en bitácoras “inmutables”, que no puedan alterarse ni borrarse, asegurando la rendición de cuentas.
Además, debe documentarse el linaje de datos (data lineage), es decir, el recorrido de cada dato desde su origen en un registro institucional, pasando por depuraciones y procesos de vinculación, hasta su utilización en un indicador.
Complementariamente, el SIIME debe publicar metadatos y reportes de calidad que incluyan tasas de vinculación y márgenes de error, y disponer de una política de

transparencia que garantice acceso público a metodologías, evaluaciones de impacto y auditorías.

3.2 Principios de seguridad

La seguridad constituye uno de los pilares del SIIME. Proteger la información sensible de las personas migrantes no solo responde a obligaciones legales, sino también a la necesidad de generar confianza institucional y social en el sistema. La experiencia internacional demuestra que los sistemas de integración de datos solo logran legitimidad cuando incorporan medidas de seguridad robustas que eviten accesos indebidos, reduzcan riesgos de divulgación y garanticen la rendición de cuentas (OECD, 2021; Stats NZ, 2017; ADR UK, 2021).

En el marco del SIIME, se proponen los siguientes principios de seguridad:

- **Perfiles de usuario diferenciados**. El acceso debe segmentarse según roles y funciones. Por ejemplo, un administrador técnico puede gestionar la infraestructura, un analista acceder únicamente a datos anonimizados para generar estadísticas, y un usuario de consulta visualizar solo indicadores agregados. Este modelo de control de accesos basado en roles (*role-based access control, RBAC*) es utilizado en el Integrated Data Infrastructure (IDI) de Nueva Zelanda y ha demostrado ser efectivo para limitar riesgos (Stats NZ, 2017).
- **Entornos seguros de acceso**. Los datos sensibles no deben circular libremente entre instituciones. En su lugar, el SIIME debe operar bajo entornos seguros de investigación (*Trusted Research Environments – TREs*), donde el acceso se realice en espacios monitoreados, sin posibilidad de descargar microdatos. Este enfoque, empleado por ADR UK y recomendado en los diseños de DARE UK, asegura que el análisis de datos se realice bajo condiciones controladas (ADR UK, 2021; DARE UK, 2023).
- **Anonimización y seudonimización**. Antes de ser compartidos con usuarios, los datos deben pasar por procesos de desidentificación. Esto implica eliminar identificadores directos (como RUN, pasaporte o dirección exacta) y reemplazarlos por códigos únicos temporales, que permiten el análisis sin exponer identidades individuales. Según la OCDE (2021), estas técnicas reducen la probabilidad de reidentificación y son parte esencial de la ética de datos en el sector público.
- **Terceros de confianza**: un organismo técnico neutral, como el INE o el SERMIG, debe actuar como custodio del sistema. Su rol es recibir los datos identificados, realizar los procesos de vinculación (determinística o probabilística) y entregar a los usuarios finales únicamente datasets anonimizados. Esta figura de "*trusted third party*" ha sido ampliamente documentada en los casos de ADR UK y en el IDI de

Nueva Zelanda, asegurando la neutralidad técnica y la protección de los datos sensibles (ADR UK, 2021; Stats NZ, 2017).

- **Protocolos de auditoría y trazabilidad**: todo acceso o modificación debe quedar registrado en sistemas de auditoría. Esto implica mantener bitácoras inmutables que documenten qué institución accedió, qué variables consultó y qué transformaciones aplicó. Adicionalmente, se debe documentar el linaje de los datos[1] (data lineage), permitiendo reconstruir cómo un registro pasó de su origen a convertirse en un indicador estadístico. Este principio, enfatizado en la guía de la UNECE (2011), es clave para la rendición de cuentas y la confianza social en el sistema.

[1] En el ámbito de gestión de datos, "lineaje de datos" (*data lineage*) se refiere a la trazabilidad completa de un dato a lo largo de su ciclo de vida: desde su origen, pasando por todas las transformaciones, integraciones y movimientos que sufre, hasta su uso final en reportes, análisis o decisiones.

Recuadro 2: Metodología de los "Five Safes" empleada Reino Unido y Nueva Zelanda.

---------- Ejemplo internacional: metodología “Five Safes” ----------

Metodología “Five Safes” para el acceso seguro a datos

El modelo Five Safes es un marco utilizado internacionalmente para garantizar el uso responsable de datos sensibles. Propone cinco dimensiones de seguridad que, en conjunto, permiten equilibrar el acceso a la información con la protección de la privacidad.

1. Personas seguras (Safe People)

Los investigadores deben ser acreditados y comprometerse a un uso responsable de los datos. Para ello asisten a capacitaciones en confidencialidad, firman compromisos legales y demuestran competencias técnicas. Quienes incumplen las normas pueden ser suspendidos o sancionados.

2. Proyectos seguros (Safe Projects)

Solo se aprueban investigaciones que aporten al interés público. No se autoriza el uso de datos para decisiones sobre casos individuales, sino para estudios que generen beneficios colectivos.

3. Entornos seguros (Safe Settings)

El acceso se realiza únicamente en entornos virtuales controlados (*Trusted Research Environments*), con múltiples capas de seguridad. Los servidores no están conectados a internet ni a redes externas, y no permiten descargas en dispositivos físicos.

4. Datos seguros (Safe Data)

Los investigadores acceden únicamente a la información estrictamente necesaria. Los datos se entregan anonimizados: se eliminan o encriptan identificadores personales como nombres, fechas de nacimiento o direcciones.

5. Resultados seguros (Safe Outputs)

Los resultados de cada proyecto son revisados antes de su publicación para asegurar que no contengan información que permita identificar a personas. Los investigadores deben aplicar técnicas de confidencialización y cumplir protocolos de validación.

3.3 Gobernanza y roles institucionales

El éxito del SIIME dependerá no solo de las tecnologías empleadas, sino de la existencia de un esquema de gobernanza robusto, que asegure coordinación interinstitucional, neutralidad técnica y responsabilidad compartida en el uso de los datos. Según la experiencia internacional, los sistemas de integración de datos más sostenibles son aquellos que combinan liderazgo político, responsabilidad técnica independiente y mecanismos de participación de las instituciones proveedoras.

En este punto es fundamental reconocer los hallazgos recientes de UNICEF (2023) "*Diagnóstico Y Recomendaciones Para Integrar Registros Administrativos Relacionados Con La Niñez.*", que advierten sobre tres debilidades recurrentes en la región respecto de las oficinas nacionales de estadísticas (ONE):

(i) una comprensión limitada, por parte de ministerios y entidades gubernamentales, de las ventajas de la vinculación e integración de registros administrativos con fines estadísticos;
(ii) la tendencia de las ONE a concebir a los ministerios solamente como proveedores de datos, y no como usuarios o socios estratégicos; y
(iii) la necesidad de que las ONE fortalezcan sus marcos de gobernanza, incluyendo mecanismos de gestión de riesgos en privacidad y seguridad. Estos elementos deben ser tomados en cuenta en el diseño del SIIME, para evitar reproducir problemas de coordinación y asegurar un modelo de gobernanza más inclusivo y colaborativo.

En esta línea, se recomienda que el SIIME se organice en torno a tres instancias principales:

- **Comité interinstitucional de datos migratorios**. Integrado por representantes de las principales instituciones proveedoras (ej. SERMIG, PDI, MINEDUC, MINSAL, SRCeI) y usuarias de la información (ej. ministerios sectoriales, academia).

  Su función es definir lineamientos estratégicos, aprobar protocolos de intercambio y supervisar la calidad y seguridad del sistema. Este modelo de comité directivo ha sido exitoso en experiencias como el ADR Network del Reino Unido, donde las decisiones sobre el acceso y uso de datos se adoptan de manera conjunta (ADR UK, 2021).

- **Unidad técnica de integración de datos**. Constituida por equipos especializados en interoperabilidad, vinculación y seguridad de la información. Esta unidad es responsable de implementar los protocolos de estandarización, gestionar la interoperabilidad entre sistemas y monitorear la calidad de las fuentes.

- **Tercero de confianza (*Trust Center*)**. Corresponde a una institución técnica neutral (por ejemplo, el INE o el propio SERMIG), que actúe como custodio central del sistema. Su rol es recibir los datos identificados, realizar los procesos de vinculación (determinísticos y/o probabilísticos), aplicar mecanismos de anonimización y entregar a los usuarios únicamente datasets protegidos. El modelo de "*trusted data custodian*" se utiliza en los entornos seguros del Reino Unido (Trusted Research Environments, TREs) y ha demostrado ser clave para mantener el equilibrio entre acceso y protección (DARE UK, 2023; ADR UK, 2021).

4. Marco Técnico

El diseño técnico del SIIME debe garantizar la interoperabilidad entre instituciones, la protección de datos personales y la posibilidad de vincular fuentes de información diversas bajo criterios consistentes y trazables. Para ello, se propone estructurar el sistema en torno a un dataset maestro, que actúe como punto de referencia y validación: el RNE

4.1 El Registro Nacional de Extranjeros como columna vertebral

El RNE constituye la base estructural del SIIME, al reunir los antecedentes administrativos de todas las personas extranjeras que han realizado gestiones ante las autoridades migratorias (SERMIG y PDI). Al centralizar información clave como RUN de extranjeros, número de pasaporte, nacionalidad, sexo, fecha de nacimiento y estatus migratorio, el RNE se convierte en la fuente de referencia primaria para cualquier proceso de integración de datos dentro del sistema.

De acuerdo con la Ley N°21.325 (2021), el RNE debe cumplir con múltiples funciones: identificar a los extranjeros en el país, registrar sus categorías migratorias, permisos y prohibiciones de ingreso, así como almacenar información sobre ingresos y egresos, solicitudes denegadas e infracciones. En la interpretación del SERMIG, esto implica que el registro debe incluir a todas las personas extranjeras que ingresen al territorio nacional, independientemente de la vía de ingreso o del carácter de su permanencia.

Para ello, el RNE debería integrar información proveniente de diversas instituciones, tales como la PDI (viajes y partes policiales), el Ministerio de Relaciones Exteriores (permisos consulares), el Ministerio Público (infracciones) y el Registro Civil (asignación de RUN y cedulación).

En este marco, el RNE está llamado a cumplir tres funciones estratégicas para el SIIME:

1. Validación y depuración de registros administrativos, permitiendo a instituciones sectoriales contrastar sus datos con una fuente oficial y reducir duplicidades.
2. Llave de enlace para la integración de datos, sirviendo como identificador maestro en vinculación determinística (ej. con RUN o pasaporte) y como referencia para técnicas probabilísticas cuando faltan identificadores únicos.
3. Soporte para la construcción de estadísticas e indicadores, facilitando la integración intersectorial y la generación de información confiable sobre empleo, salud, educación, vivienda e inclusión social de la población migrante.

Estado actual del RNE (2025): desafíos y limitaciones

Si bien el RNE está concebido legalmente como la columna vertebral del SIIME, en la práctica su desarrollo institucional y tecnológico aún enfrenta limitaciones importantes. Desde 2022, el SERMIG trabaja en paralelo en dos frentes:

- Interoperabilidad interna, que busca integrar sus propios sistemas fragmentados: el histórico “B3000”, basado en expedientes personales, y el sistema “Simple”, orientado a la gestión de trámites desde 2020.
- Interoperabilidad externa, que persigue articular los datos de la PDI, SRCeI, MINREL y Ministerio Público, mediante cargas externas o futuros servicios de interoperabilidad.

En 2025, la infraestructura del RNE sigue en fase de diseño. Si bien existe un proyecto de desarrollo de software (“RNE software”) que integraría estas fuentes en una única interfaz, éste fue paralizado en 2022 y no ha retomado una hoja de ruta clara. Mientras tanto, la base de datos B3000 continúa siendo utilizada como referencia de facto para la integración de información interna y externa.

Este escenario implica que, aunque el RNE tiene el mandato legal y el potencial técnico para ser la columna vertebral del SIIME, en su estado actual no cumple plenamente con las condiciones de interoperabilidad y consulta necesarias. La coexistencia de sistemas fragmentados y la ausencia de un repositorio consolidado dificultan su uso como registro maestro en el corto plazo.

En consecuencia, el diseño del SIIME deberá considerar un enfoque gradual y adaptativo, acompañando la consolidación del RNE y estableciendo mecanismos transitorios que permitan avanzar en la integración de datos mientras se fortalecen las capacidades tecnológicas e institucionales del SERMIG.

4.2 Metodologías de vinculación de datos

El proceso de integración de registros en el SIIME requiere metodologías que permitan identificar, con el mayor grado de certeza posible, si dos registros de distintas fuentes corresponden a la misma persona. La experiencia internacional muestra que no existe un único método aplicable a todos los contextos, sino que es necesario combinar enfoques según la disponibilidad y calidad de los identificadores (UNECE, 2011).

En términos generales, la vinculación puede realizarse mediante dos enfoques principales: determinístico y probabilístico, que no son excluyentes, sino complementarios.

4.2.1 Vinculación determinística

Este enfoque establece coincidencias exactas entre registros. Su forma más sólida es la vinculación con identificadores únicos como el RUN, que en Chile constituye la clave primaria para procesos de alta precisión y bajo costo computacional.

No obstante, muchos registros que contienen información sobre población migrante, especialmente vinculados a situaciones irregulares, carecen de un identificador único confiable. En estos casos se recurre a reglas determinísticas basadas en múltiples variables, como nombre completo, sexo, fecha de nacimiento o nacionalidad. Esta aproximación es más frágil, pero puede dar resultados útiles si se acompañan de protocolos claros de calibración, auditorías regulares y revisión manual de casos dudosos.

En ausencia de identificadores únicos, la vinculación determinística tiene ventajas: es relativamente sencilla de implementar, menos costosa en términos computacionales y altamente auditable, ya que las reglas aplicadas son explícitas. Sin embargo, también presenta limitaciones: es muy sensible a errores tipográficos o de digitación, ofrece poca flexibilidad frente a datos incompletos o inconsistentes, y puede excluir a personas legítimas si las variables no coinciden exactamente (ADR UK, 2021).

4.2.2 Vinculación probabilística

La vinculación probabilística se utiliza cuando no se dispone de un identificador único confiable en todas las fuentes. A diferencia del enfoque determinístico, que exige coincidencias exactas, este método asigna “pesos de coincidencia” (“*match weights*”) a los pares de registros según el patrón de similitudes y discrepancias en las variables comparadas (ej. nombre, apellidos, sexo, fecha de nacimiento, nacionalidad).

Un error común en la literatura es afirmar que la vinculación probabilística estima directamente la probabilidad de que dos registros correspondan a la misma persona. Según

Harron et al. (2020) en *Demystifying probabilistic linkage: Common myths and misconceptions* esto constituye un mito *(*En el Anexo A se encuentra una tabla resumen sobre los mitos y verdades de la vinculación probabilística).

En realidad, los algoritmos generan puntajes que se correlacionan con la probabilidad de ser un match, pero no representan probabilidades exactas. Los pesos aumentan cuando la probabilidad de coincidencia es alta y disminuyen cuando es baja, pero dependen de supuestos (como la independencia de los valores m y u) que no siempre se cumplen. Por ello, los "*match weights*" deben entenderse como indicadores relativos de similitud, y no como probabilidades absolutas de coincidencia.

El modelo clásico de referencia es el de Fellegi y Sunter (1969), que utiliza estos pesos para clasificar pares de registros en tres categorías: *match seguro*, *no match* y *posible match* (revisión manual). Este enfoque ha sido adoptado en numerosos sistemas estadísticos oficiales, especialmente en países que carecen de identificadores únicos universales.

En ausencia de identificadores únicos, la vinculación probabilística ofrece ventajas claras: permite manejar errores de digitación y variantes ortográficas, trabajar con datos incompletos y aumentar la cobertura de la integración. Sin embargo, también tiene desafíos: es metodológicamente más compleja, requiere mayor capacidad computacional y sus resultados dependen de umbrales probabilísticos que deben ser cuidadosamente documentados y auditados, lo que puede afectar su interpretabilidad.

Una innovación reciente es el desarrollo de Splink, un paquete de Python de código abierto desarrollado inicialmente por el Ministerio de Justicia en el Reino Unido y adoptado por la Oficina Nacional de Estadísticas. Splink permite implementar modelos probabilísticos a gran escala, aplicando algoritmos de *expectation–maximization*[2] y técnicas de *blocking*[3] para reducir el número de comparaciones necesarias. Gracias a estas optimizaciones, es posible procesar bases masivas con un rendimiento mucho mayor al de herramientas tradicionales.

La tabla 2 continuación muestra un resumen y comparación de ambos métodos de vinculación y deduplicación de datos.

[2] **Expectation-maximizaton**: Algoritmo estadístico iterativo que permite estimar parámetros de modelos cuando existen variables no observadas o incompletas. En vinculación de registros, el EM se utiliza para calcular las probabilidades de acuerdo y desacuerdo entre variables (valores *m* y *u*), actualizando los parámetros hasta que convergen a un valor estable (Dempster, Laird & Rubin, 1977).

[3] **Blocking**: Algoritmo estadístico iterativo que permite estimar parámetros de modelos cuando existen variables no observadas o incompletas. En vinculación de registros, el EM se utiliza para calcular las probabilidades de acuerdo y desacuerdo entre variables (valores *m* y *u*), actualizando los parámetros hasta que convergen a un valor estable (Dempster, Laird & Rubin, 1977).

Tabla 2: Comparación de los métodos de vinculación de registros

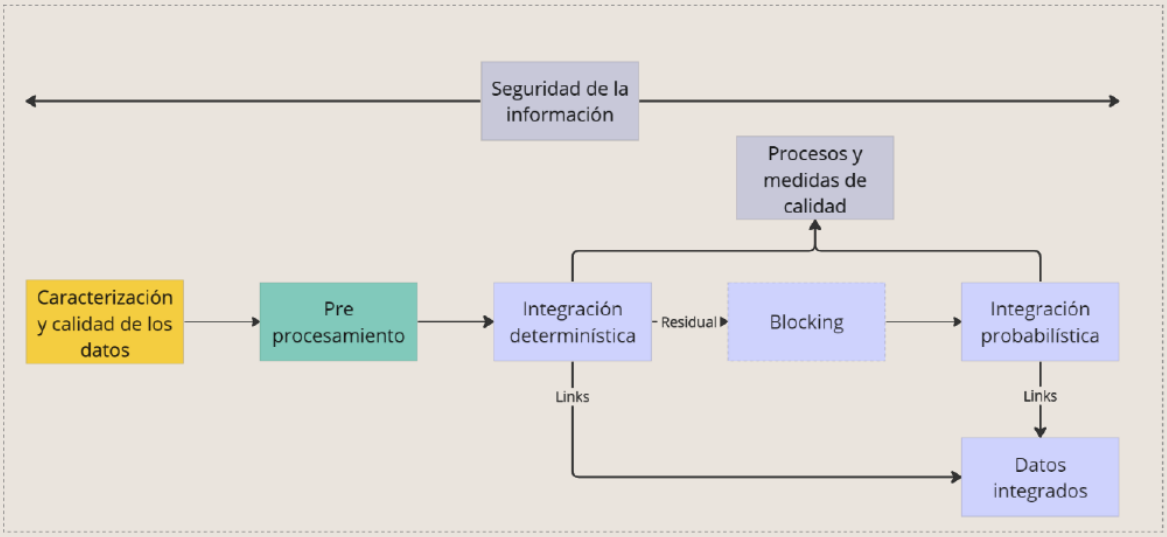
ASPECTO	VINCULACIÓN DETERMINÍSTICA	VINCULACIÓN PROBABILÍSTICA
DEFINICIÓN	Coincidencia exacta de identificadores o combinaciones de variables bajo reglas explícitas.	Asigna pesos de coincidencia a variables y calcula puntajes que estiman la probabilidad relativa de ser un match.
VENTAJAS	- Fácil de implementar y comprender. - Bajo costo computacional. - Transparencia en reglas aplicadas. - Fácil auditoría.	- Flexible ante errores tipográficos, variantes ortográficas y datos faltantes. - Mayor cobertura en bases heterogéneas. - Permite combinar múltiples variables con distinto peso. - Puede generar mejores resultados cuando no hay identificadores únicos.
DESVENTAJAS	- Muy sensible a errores de digitación y datos incompletos. - Poca flexibilidad: si no hay coincidencia exacta, no hay match. - Riesgo de exclusión de casos legítimos.	- Mayor complejidad metodológica. - Requiere mayor capacidad computacional. - Resultados dependen de supuestos estadísticos (ej. independencia de variables). - Menor interpretabilidad: requiere documentación clara y auditorías.
CASOS DE USO ÓPTIMOS	Cuando existen identificadores únicos robustos (ej. RUN, pasaporte) o variables estandarizadas y de alta calidad.	Cuando no existen identificadores únicos, o los registros contienen errores, inconsistencias o datos faltantes.

Recomendación sobre la vinculación de registros en el SIIME

En el SIIME, la vinculación determinística con RUN debe ser la estrategia preferida por su precisión y eficiencia. Sin embargo, se debe prever que una parte significativa de los registros migratorios no contará con este identificador. En esos casos, la vinculación determinística con múltiples variables puede ser un recurso inicial de bajo costo, aunque limitado frente a la heterogeneidad de los datos. Para escenarios más complejos o donde se requiera mayor cobertura, la vinculación probabilística, apoyada en herramientas modernas como Splink, ofrece soluciones más robustas, aunque con mayores requerimientos técnicos y metodológicos.

En todos los casos, resulta indispensable que los procesos de vinculación se realicen bajo protocolos claros, auditorías de calidad y documentación pública de los márgenes de error, siguiendo las recomendaciones internacionales y las prácticas consolidadas en Stats NZ (2017) y ADR UK (2021).

Figura 1: Propuesta de diseño de un proceso de vinculación de registros en el SIIME



Fuente: Elaboración propia con insumos de la conferencia ADR UK 2023 (GLADIS Pipeline, ONS 2023 https://ijpds.org/article/view/2219)

4.3 Arquitectura técnica

El SIIME debe concebirse como una arquitectura modular, escalable y lo suficientemente flexible para incorporar nuevas fuentes de datos a lo largo del tiempo, garantizando siempre altos niveles de seguridad, privacidad y calidad. Este diseño debe apoyarse en principios reconocidos de interoperabilidad de datos y en experiencias internacionales que han demostrado su efectividad, tales como ADR UK, DARE UK y los lineamientos internacionales de interoperabilidad estadística.

A continuación, se describen los modelos principales, sus ventajas, desventajas y las consideraciones técnicas clave, incluidas referencias concretas:

4.3.1 Modelos de arquitectura posibles

a) Intercambio descentralizado (Arquitectura federada)

En este modelo, cada institución mantiene la posesión y control de su propia base de datos, exponiendo únicamente la información necesaria a través de interfaces seguras y estandarizadas, como APIs o servicios web. El SIIME actuaría como un "bus de interoperabilidad" que se apoya en un identificador maestro, como el RUN o el número de extranjero del RNE para alinear registros y permitir validaciones en tiempo real. Este modelo ofrece la ventaja de reducir la concentración de datos sensibles y preservar la autonomía institucional, además de favorecer actualizaciones cercanas al tiempo real y reducir la duplicación en el almacenamiento.

Sin embargo, implica desafíos técnicos y organizativos considerables: depende de la madurez de cada institución para mantener interfaces operativas, requiere estándares comunes de metadatos y semántica muy rigurosos, y puede sufrir problemas de latencia si uno de los sistemas no está disponible. Por otro lado, también está la complejidad técnica de al implementar autenticación y autorización compartidas.

La experiencia del *Federated Architecture Blueprint* de DARE UK es ilustrativa: ha mostrado la viabilidad de consultas distribuidas desde entornos seguros de investigación (*TRE*), aunque con elevados costos de coordinación y gobernanza técnica.

b) Repositorio centralizado de integración

En este enfoque, las instituciones envían extractos periódicos de sus datos a un repositorio común gestionado por un ente de confianza, como podría ser el INE o el SERMIG. En este repositorio se realizan los procesos de integración, depuración, vinculación y análisis, lo que permite generar estadísticas consolidadas y estudios longitudinales.

Este modelo otorga un control más directo sobre la calidad, la estandarización y la seguridad, y facilita la creación de indicadores y productos estadísticos nacionales. No

obstante, concentra los riesgos de seguridad al centralizar datos sensibles, exige protocolos muy sólidos de anonimización y acceso, y requiere inversiones relevantes en almacenamiento y procesamiento. Por otro lado, la actualización puede ser menos frecuente, lo que puede afectar la oportunidad de los datos.

El caso chileno del Registro Estadístico de Población (REP) constituye un ejemplo de este enfoque, y en el plano internacional destaca la *Integrated Data Infrastructure* (IDI) de Stats NZ, reconocida como buena práctica en la gestión integrada de datos administrativos.

c) Modelo híbrido (recomendado)

La evidencia internacional muestra que ninguno de los modelos anteriores resulta suficiente de manera aislada. Por ello, la tendencia más consolidada es avanzar hacia esquemas híbridos, que combinan los beneficios de ambos. En un modelo híbrido, las validaciones operativas y las necesidades de gestión se realizan a través de mecanismos federados en tiempo real, mientras que los productos estadísticos y los análisis longitudinales se generan a partir de un repositorio central consolidado bajo estricta gobernanza.

Este enfoque ha sido adoptado por ADR UK, que combina repositorios centralizados con mecanismos de acceso federado mediante "*Safe Pods*". Para el caso chileno, un modelo híbrido permitiría aprovechar la capacidad de validación inmediata con el RNE para las operaciones administrativas, al mismo tiempo que consolidaría un repositorio central de uso estadístico bajo la coordinación del INE y el SERMIG.

4.3.2 Adaptación al caso chileno

La definición de la arquitectura del SIIME debe considerar el contexto institucional y tecnológico específico de Chile. En primer lugar, el país dispone de un identificador único robusto, complementado por el RNE que actualmente se encuentra en proceso de consolidación. Estos instrumentos constituyen una base sólida para articular tanto la interoperabilidad federada como la integración centralizada. A diferencia de otros países donde la ausencia de un identificador transversal obliga a depender intensamente de métodos probabilísticos, en Chile es posible avanzar de manera más eficiente hacia esquemas híbridos con un alto grado de certeza en la vinculación de registros.

Un segundo elemento relevante es la experiencia acumulada por el REP, que demuestra la factibilidad de construir un repositorio central de referencia a partir de múltiples fuentes administrativas. El REP ha evidenciado tanto el potencial del RUN como llave de enlace, como los desafíos de calidad en ciertos grupos poblacionales. Estas lecciones resultan esenciales para orientar el diseño del SIIME y evitar la repetición de problemas ya identificados.

Asimismo, el marco institucional chileno ofrece ventajas y desafíos particulares. Por un lado, la existencia de un organismo técnico como el INE y de una autoridad especializada en migraciones como el SERMIG permite imaginar un modelo de gobernanza compartida donde ambas instituciones ejerzan funciones diferenciadas: el INE como garante de calidad estadística y el SERMIG como responsable de la gestión operativa y de la retroalimentación administrativa. Por otro lado, la diversidad de sistemas sectoriales y su heterogeneidad tecnológica obligan a una inversión importante en estándares comunes de interoperabilidad, metadatos y mecanismos de seguridad de la información.

Finalmente, la adaptación chilena de la arquitectura técnica debe también proyectarse en relación con el futuro Sistema Nacional de Gestión de Datos, actualmente en discusión legislativa (Ministerio de Hacienda, 2025). De aprobarse, este sistema establecería un marco nacional de interoperabilidad y gestión de datos que podría convertirse en un soporte clave para el SIIME, al reducir costos de coordinación, elevar la eficiencia del intercambio interinstitucional y garantizar estándares unificados de gobernanza. La articulación entre el SIIME y este futuro sistema debe ser planificada desde el inicio, de modo que el primero se integre como una aplicación sectorial especializada dentro de un marco nacional más amplio.

En términos prácticos, la adaptación chilena del modelo híbrido debería seguir una lógica gradual. En una fase inicial, se recomienda priorizar la construcción de repositorios centralizados de menor escala, por ejemplo, en ámbitos de salud y educación, que sirvan como pilotos para probar rutinas de vinculación, protocolos de anonimización y métricas de calidad. Paralelamente, se deben desarrollar las interfaces federadas que permitan validar en tiempo real información crítica, como la identidad y situación administrativa de una persona en el RNE. Finalmente, en una fase de madurez, el SIIME debería consolidarse como una infraestructura nacional estable, con un repositorio central actualizado periódicamente y mecanismos federados que garanticen la pertinencia de la información en la gestión cotidiana.

En resumen, la arquitectura técnica del SIIME en Chile debe ser híbrida, con una columna vertebral basada en el RNE y el RUN, un repositorio central bajo gobernanza robusta para fines estadísticos e investigación, y una capa federada para validaciones administrativas.

5. Casos prácticos

5.1 Registro Estadístico de Población (INE)[4]

El Registro Estadístico de Población (REP), un proyecto del Instituto Nacional de Estadísticas (INE), es un ejemplo concreto de cómo es posible integrar distintas fuentes administrativas para construir un sistema robusto de conteo, localización y caracterización básica de la población residente (nacional y extranjera).

El objetivo del REP es consolidar datos de población desagregados, oportunos y de calidad, reduciendo los costos de operativos de terreno (como censos o encuestas) y la carga sobre informantes y encuestadores. A mediano plazo, constituye un primer paso hacia la implementación de censos mixtos o incluso censos basados únicamente en registros administrativos.

Fuentes de datos utilizadas

En 2024, el REP se construye a partir de tres grandes registros:

- **Registro Civil e Identificación (SRCeI)**: contiene todos los RUN asignados a personas chilenas y extranjeras. (≈26 millones de registros). *Columna vertebral del REP*.
- **FONASA:** registro mensual de beneficiarios del sistema público de salud. (≈17 millones de registros).
- **SUSESO:** registro mensual de trabajadores protegidos por el seguro de accidentes laborales. (≈8 millones de registros).

La vinculación se realiza utilizando como llave principal el RUN, lo que permite determinar en qué registros está presente cada persona (solo en SRCeI, en SRCeI+ FONASA o SUSESO, o en los tres simultáneamente).

Principales procesos de vinculación de datos

El proceso de vinculación de registros se organiza en dos grandes etapas:

[4] Es importante detallar que el REP es un producto en proceso de contrucción, por lo que toda la información presentada en este informe está sujeta a cambios y modificaciones. La descripción aquí expuesta representa el estado actual del proyecto y no necesariamente lo que será publicado por los canals formales del INE.

1. Deduplicación

Todas las bases contienen personas con más de un registro, por lo que es necesario depurar y conservar un único registro por persona.

- *Determinística*: busca coincidencia exacta en variables clave (nombre completo y fecha de nacimiento en SRCeI y FONASA; nombre completo y sexo en SUSESO). Por ejemplo, en el Registro Civil se identificaron ≈38.000 duplicados entre 26 millones de registros.
- *Probabilística*: se emplean algoritmos de vinculación que calculan un peso de coincidencia en base a múltiples variables sociodemográficas (nombres, apellidos, fecha de nacimiento, nacionalidad, región de residencia). El INE utiliza el software Splink en Python, que permite realizar resolución de entidades sin identificadores únicos, reduciendo significativamente los tiempos de cómputo al comparar solo pares candidatos y no todos los registros posibles.

2. Vinculación

Una vez depuradas las bases, la vinculación se realiza de manera determinística a través del RUN, conectando las tres fuentes (SRCeI, FONASA y SUSESO). Esto permite alcanzar un alto porcentaje de correspondencias, lo que evidencia la solidez del RUN como identificador único para la integración de bases de datos. No obstante, un desafío a corto plazo consiste en probar Splink para vincular aquellos registros que no logran un match directo.

Sobre la base ya vinculada, se aplica un proceso adicional de deduplicación determinística utilizando como criterios las variables: nombre, primer apellido, segundo apellido y fecha de nacimiento. Este paso permite identificar, por ejemplo, a personas que figuran con un NIP/NIC en FONASA y que además poseen un RUN otorgado por el SRCeI.

La base de datos final se organiza en tres tablas:

- Tabla de hechos: contiene todos los identificadores (RUN o, en su defecto, NIP/NIC) presentes en las tres bases durante 2024.
- Tabla demográfica: incluye las variables sociodemográficas (nombres, apellidos, fecha de nacimiento, sexo, nacionalidad y estado civil).
- Tabla geográfica: concentra la información territorial (región y/o comuna).

Para decidir qué variables conservar de cada fuente, se definieron criterios de priorización basados en la calidad de los registros administrativos:

- Para la mayoría de las variables demográficas (nombres, apellidos, sexo y nacionalidad) se prioriza la información en el siguiente orden: SRCeI, FONASA y SUSESO.
- En el caso de las variables geográficas, se prioriza primero la información de FONASA por sobre la del SRCeI.
- Las variables de señales de vida (una variable binaria que indica si la persona estuvo presente en los registros por al menos seis meses consecutivos) se obtienen a partir de FONASA y SUSESO.

Consideraciones tecnológicas

Las técnicas probabilísticas, aunque más precisas, son intensivas en recursos computacionales. Gracias a Splink y a reglas determinísticas previas, se redujo el universo de comparación en el caso del SRCeI de 26 millones de posibles pares a ≈1,6 millones, optimizando tiempo y memoria. Aun así, el procesamiento requiere servidores de alta capacidad (≈190 GB de RAM, equivalentes a 30 procesadores core i9), siendo estas las características de los servidores disponibles y el procesamiento con Splink nunca excede esta capacidad.

Consideraciones institucionales

Actualmente, el intercambio de datos entre el INE y las instituciones proveedoras es limitado en cuanto a interoperabilidad y transparencia de procesos. Las instituciones entregan anualmente archivos planos de gran tamaño, sin documentación detallada sobre preprocesamientos o depuraciones realizadas.

El escenario ideal sería avanzar hacia un modelo de interoperabilidad en tiempo real, aprovechando mecanismos como los *webs services* del SRCeI, lo que permitiría trabajar con datos “en bruto” y mejorar la trazabilidad y calidad del REP. Sin embargo, hoy en día los convenios de intercambio de datos condicionan en gran medida las posibilidades técnicas y legales del proceso.

---------- Lecciones del REP para el SIIME ----------

A partir de la experiencia del REP se identifican aprendizajes clave que pueden orientar el desarrollo del SIIME:

- Volumen y capacidad tecnológica
  La magnitud de los registros (decenas de millones de filas) exige procesos intensivos en cómputo. Es imprescindible planificar un uso eficiente de tecnología, optimización de algoritmos y contar con infraestructura de procesamiento de alto rendimiento.
- Manejo de duplicados requiere criterio experto
  Casos como FONASA muestran que no basta con reglas simples de deduplicación. Es necesario calibrar criterios según el balance entre errores tipo I (falsos positivos) y tipo II (falsos negativos), incorporando la mirada de expertos del negocio para definir reglas de emparejamiento adecuadas.
- Calidad heterogénea entre registros
  Algunas fuentes, como SUSESO, presentan inconsistencias frecuentes (ej. variación en nacionalidad y sexo a nivel individual). Esto refuerza la necesidad de contar con protocolos de calidad y armonización de variables antes de la integración.
- RUN como identificador robusto
  El RUN se confirma como la llave más sólida para la integración de registros, dado el alto porcentaje de emparejamiento que permite alcanzar.
- Brechas de cobertura poblacional
  Al contrastar el REP con proyecciones de INE y CELADE se identifican vacíos en ciertos grupos:
  - Subregistro de defunciones → Sobreestima a la población de 90 y más años.
  - Subregistro en el tramo 0–10 años.
  - Subcobertura de población en edad de trabajar vinculada al sector informal.
    Esto sugiere la necesidad de flexibilizar criterios de "señal de vida" y ajustar metodologías para mejorar la representación estos grupos (por ejemplo, la población extranjera irregular).
- Interoperabilidad y gobernanza institucional
  El trabajo actual con archivos planos limita la trazabilidad y calidad de los datos. Avanzar hacia interoperabilidad en tiempo real (web services, APIs seguras) y establecer convenios claros sobre roles, responsabilidades y estándares de calidad será determinante para la sostenibilidad del SIIME.

5.2 Estimación de Personas Extranjeras Residentes en Chile (EPE)

Un caso concreto de cómo se vinculan registros administrativos en Chile es la Estimación de Personas Extranjeras (EPE) 2023, elaborada por el Instituto Nacional de Estadísticas (INE) en conjunto con el Servicio Nacional de Migraciones (SERMIG) desde 2018. Esta estimación busca actualizar anualmente el stock de población extranjera residente en Chile al 31 de diciembre de cada año, complementando la base del Censo 2017 con registros administrativos provenientes de distintas instituciones públicas.

La experiencia de la EPE resulta especialmente relevante para el diseño del SIIME, pues constituye un caso pionero en la integración de fuentes administrativas de migración, educación, salud y seguridad, mediante procesos de depuración y vinculación determinística.

Fuentes de datos utilizadas

La componente de registros administrativos de la EPE se nutre de varias fuentes que cubren tanto a la población extranjera regular (con permisos y trámites registrados) como a la irregular (sin documentación migratoria). Las principales son:

- Servicio Nacional de Migraciones (SERMIG): registros de permisos de residencia, refugio, sanciones administrativas y prórrogas de turismo.
- Ministerio de Educación (MINEDUC): registro de matrícula escolar, incluyendo estudiantes que no cuentan con RUN y reciben un Identificador Provisorio Escolar (IPE).
- Policía de Investigaciones (PDI): registros de control de fronteras (entradas y salidas) y partes policiales asociados a ingresos o permanencias irregulares.
- Ministerio de Relaciones Exteriores (MINREL): permisos y visas consulares otorgadas fuera del país.
- Servicio de Registro Civil e Identificación (SRCeI): registros de defunciones de personas extranjeras y actualizaciones de RUN.
- Empadronamiento biométrico (2023–2024): registro especial de migrantes en situación irregular mayores de 18 años.

Cada fuente aporta un componente distinto de la población migrante, lo que exige armonización previa y deduplicación de registros antes de consolidar la base final.

Procesos principales de vinculación de registros

El proceso metodológico de la EPE se organiza en dos grandes etapas: (1) deduplicación de registros internos y (2) integración interinstitucional.

1. Deduplicación de registros internos

Cada base contiene casos de una misma persona registrada más de una vez. Para resolverlo, se aplican procesos de deduplicación basada en reglas determinísticas: se utilizan variables únicas como RUN, número de pasaporte o una serie de reglas con combinaciones exactas de variables como nombres, apellidos y fecha de nacimiento.

Ejemplo:MINEDUC
La base de matrícula escolar contiene estudiantes con más de un RUN ALU o con IPE. Para depurarla, se implementa un algoritmo con 108 combinaciones de variables clave (nombres, apellidos, fecha de nacimiento, sexo, establecimiento–RBD). Este nivel de complejidad permite identificar registros duplicados y conservar solo un registro consolidado por estudiante.

2. Integración interinstitucional

Una vez depuradas las bases individuales, se procede a integrarlas. La vinculación entre instituciones se realiza también con reglas determinísticas, pero adaptadas a cada cruce. El proceso contempla cinco etapas:

- **Etapa 1**: **Consolidación Sermig–Mineduc**.
  Se cruzan los registros históricos de SERMIG con la matrícula escolar de MINEDUC. Para maximizar la vinculación, se utilizan 106 combinaciones de variables (RUN, nombres, apellidos, fecha de nacimiento, sexo, nacionalidad).
- **Etapa 2**: **Consolidación Sermig–Minrel**.
  Se integran los registros de visas consulares otorgadas por MINREL entre 2017 y 2023, junto con las visas entregadas directamente por SERMIG a partir de mayo de 2022 (tras la entrada en vigencia de la Ley 21.325). Esto permite capturar residencias otorgadas fuera de Chile, ampliando la cobertura del stock migratorio.
- **Etapa 3**: **Unión de bases**.
  Se combinan los resultados de las etapas 1 y 2 en una base consolidada de personas extranjeras residentes, tanto dentro de Chile como con visas consulares previas a su ingreso.
- **Etapa 4**: **Vinculación con PDI (entradas y salidas)**.
  Una etapa crítica es el cruce de los registros de personas con el último movimiento de frontera registrado en las bases de datos de la PDI, que permiten actualizar el stock descontando las salidas definitivas.

- El procesamiento se realiza en SQL por la PDI, preparando previamente las bases por el Sermig.
- Variables utilizadas: RUN, nombres, apellidos, fecha de nacimiento, nacionalidad, número de pasaporte, y rangos de fechas (desde/hasta).
- Para aumentar la cobertura, se crean llaves alternativas como *PATMATNOM* (apellido paterno + materno + nombre), *PATNOM* (apellido paterno + nombre) y fecha de nacimiento (*FECNAC*).
- No existe un número fijo de llaves: se prueban distintas combinaciones para maximizar la tasa de vinculación *persona–viaje*.

- **Etapa 5**: **Vinculación con defunciones (SRCel)**.
  Finalmente, se integran los registros de defunciones de extranjeros (15.163 entre 2015 y 2023). Tras estandarizar nombres, apellidos y RUN, se identificaron 3.344 defunciones en 2023, de las cuales 3.328 corresponden al componente regular y 166 al irregular.

Resultados de la EPE 2023

El ejercicio permitió estimar un total de 1,9 millones de personas extranjeras residentes en Chile al 31 de diciembre de 2023:

- 1,58 millones en situación regular.
- 337 mil en situación irregular, estimados principalmente mediante partes policiales, matrículas sin RUN y el empadronamiento biométrico.

La vinculación con movimientos de frontera PDI alcanzó niveles altos en el componente regular (90,7%), mientras que en el componente irregular (37,4%) fue más limitada, debido a la ausencia de identificadores únicos y la calidad heterogénea de las fuentes.

---------- Lecciones de la EPE para el SIIME ----------

La experiencia de la EPE entrega lecciones valiosas para el diseño del SIIME:

- RUN como identificador robusto: es la llave más confiable para la vinculación de registros. Sin embargo, en los casos sin RUN se requieren técnicas determinísticas basadas en combinaciones de variables.

- Vinculación determinística efectiva: la experiencia demuestra que es posible generar estimaciones robustas de población extranjera utilizando reglas de vinculación determinísticas sobre registros administrativos, siempre que se cuente con procesos rigurosos de depuración.

- Deduplicación previa como paso indispensable: deben definirse reglas claras y calibradas para eliminar duplicados, adaptadas a la calidad de las variables en cada fuente.

- Calidad heterogénea de las fuentes: algunas bases, como la del SRCeI, presentan buena trazabilidad, mientras que otras requieren mayores procesos de estandarización y validación. Esto resalta la importancia de establecer protocolos comunes de calidad en el SIIME.

- Indicadores de calidad: en la versión 2023 se incluye por primera vez un documento de calidad donde se calculan indicadores sobre los registros administrativos utilizados y sobre los procesos de vinculación proceso.

- Visibilización de población de difícil alcance: la EPE ha permitido cuantificar y caracterizar segmentos invisibilizados en registros tradicionales, como las personas migrantes en situación irregular. Esta experiencia demuestra que la integración de fuentes administrativas heterogéneas puede ampliar la cobertura estadística hacia grupos históricamente subregistrados.

- Colaboración interinstitucional efectiva: la EPE constituye un caso concreto de cooperación técnica entre múltiples organismos, donde cada aporta insumos, criterios y registros que, al vincularse, generaron una estimación más robusta. Para el SIIME, esta experiencia muestra que el éxito del sistema no depende solo de tecnología o metodologías, sino también de acuerdos de gobernanza y confianza entre instituciones que permitan compartir y armonizar datos bajo estándares comunes.

6. Recomendaciones y próximos pasos

El éxito del SIIME dependerá en gran medida de la claridad con que se definan los roles institucionales, la gobernanza intersectorial y los mecanismos técnicos de vinculación. Con base en los hallazgos de este informe, y en experiencias internacionales relevantes, se proponen las siguientes recomendaciones estratégicas:

1. Organismo coordinador

En primer lugar, resulta prioritario designar un organismo coordinador del SIIME. La evidencia comparada muestra que los sistemas de integración de datos requieren de un ente rector que actúe como "tercero de confianza", con atribuciones claras para garantizar estándares comunes, asegurar el cumplimiento legal y arbitrar eventuales controversias interinstitucionales. En el caso chileno, una opción plausible sería la creación de una Subcomisión de Estadísticas Migratorias conjunta entre el SERMIG y el INE, donde el primero asuma la conducción operativa y el segundo lidere la dimensión estadística y metodológica.

2. Protocolos interinstitucionales

En segundo lugar, se recomienda elaborar protocolos interinstitucionales de intercambio de datos, que establezcan estándares técnicos (formatos, metadatos, interoperabilidad semántica), mecanismos de seguridad (anonimización, seudonimización, cifrado), y reglas de acceso diferenciadas según perfiles de usuario. Dichos protocolos deberían basarse en las prácticas del marco "*Five Safes"* y en los lineamientos de interoperabilidad internacionales, garantizando un equilibrio entre utilidad analítica y protección de la privacidad.

3. Pilotos de vinculación y deduplicación)

Un tercer paso es la implementación de pilotos de deduplicación vinculación determinística y probabilística en áreas prioritarias como educación, salud y trabajo. Estos pilotos permitirán probar, en un entorno controlado, la factibilidad técnica de los procesos de vinculación, los umbrales de error aceptables y la pertinencia de distintos algoritmos.

Estos pilotos permitirán el diseño de un "pipeline" o proceso de trabajo para la deduplicación vinculación de registros administrativos ante distintos escenarios y casos de uso.

4. Fortalecimiento institucional

Paralelamente, debe establecerse un plan de fortalecimiento institucional y de desarrollo de capacidades técnicas. Las lecciones del diagnóstico de UNICEF (2023) subrayan la importancia de que el INE no sea visto únicamente como receptora de datos, sino también como un socio estratégico de los ministerios sectoriales. Esto implica generar instancias de capacitación, intercambio y apropiación institucional, que fortalezcan la confianza y fomenten una gobernanza más colaborativa.

5. Hoja de ruta escalonada

Finalmente, se sugiere definir una hoja de ruta escalonada a corto, mediano y largo plazo. En el corto plazo, el foco debe estar en pilotos de vinculación y en la definición de protocolos de intercambio. En el mediano plazo, en la consolidación de un modelo híbrido de arquitectura técnica que combine federación para operaciones administrativas y centralización para productos estadísticos. Y en el largo plazo, en la plena integración del SIIME dentro del futuro Sistema Nacional de Gestión de Datos, actualmente en discusión legislativa, lo que permitirá asegurar sostenibilidad, eficiencia y comparabilidad internacional.

En conjunto, estas recomendaciones ofrecen un marco de acción concreto para avanzar de manera ordenada y sostenible hacia la consolidación de un sistema integrado de información sobre migración y extranjería en Chile con contemple a todas las instituciones del Estado que registren y administren información sobre población extranjera.

Referencias

1. ADR UK. (2021). *ADR England Strategy 2021–2026*. UK Research and Innovation.

2. DARE UK. (2023). *Federated Architecture Blueprint – Initial Draft*. Digital Research Infrastructure for the UK.

3. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1–38.

4. Doidge, J. C., & Harron, K. (2018). *Demystifying probabilistic linkage: Common myths and misconceptions*. *International Journal of Population Data Science*, 3(1).

5. Harron, K. (2016). *Introduction to Data Linkage. Administrative Data Research Network* (ADRN).

6. Harron, K. L., [et al.]. (2017). *A guide to evaluating linkage quality for the analysis of linked data.*

7. Naciones Unidas. (2025, marzo). *Recommendations on Statistics of International Migration and Temporary Mobility*. United Nations Statistics Division (UNSD).

8. OECD. (2021). *Good Practice Principles for Data Ethics in the Public Sector*. Paris: OECD Publishing.

9. Stats NZ. (2017). *Integrated Data Infrastructure (IDI): Overarching Privacy Impact Assessment*. Statistics New Zealand.

10. UNECE. (2011). *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. Geneva: United Nations.

11. UNICEF (2023) *Diagnóstico Y Recomendaciones Para Integrar Registros Administrativos Relacionados Con La Niñez.*

12. Winkler, W. E. (2006). *Overview of record linkage and current research directions*. Research Report Series, Statistics #2006-2. U.S. Census Bureau.

Anexos

Anexo A. Elaboración a partir de Harron (2017): *A guide to evaluating linkage quality for the analysis of linked data.*

Mito	Realidad
1. La vinculación probabilística y la determinística son métodos completamente distintos.	Son fundamentalmente equivalentes: cada regla determinística equivale a un peso probabilístico, y cada umbral probabilístico corresponde a un conjunto de reglas.
2. La vinculación probabilística se basa en la probabilidad de que dos registros sean un match.	Se basa en puntajes (match weights) que se correlacionan con la probabilidad, pero no son probabilidades absolutas.
3. La vinculación probabilística es intrínsecamente imperfecta o imprecisa.	Con variables suficientes y de calidad, puede alcanzar discriminación perfecta, igual que la determinística.
4. La vinculación probabilística produce más falsos positivos, mientras que la determinística produce más falsos negativos.	En ambos casos hay un trade-off entre falsos y verdaderos matches; se controla ajustando reglas (determinística) o umbrales (probabilística).
5. La vinculación probabilística requiere revisión manual obligatoria.	La revisión manual es opcional en ambos enfoques y depende de cómo se definan umbrales o reglas.
6. Solo la vinculación probabilística permite aceptar desacuerdos entre variables.	La determinística también puede hacerlo, explícita (tolerando diferencias) o implícitamente (ignorando variables).
7. Solo la vinculación probabilística permite manejar acuerdos parciales (ej. nombres similares o fechas incompletas).	La determinística también puede, usando reglas como Soundex o similitud de cadenas; la probabilística lo hace de forma más flexible.
8. La vinculación probabilística refleja la incertidumbre de los matches, la determinística no.	En la práctica, ambos suelen tratar los matches como definitivos; técnicas avanzadas como la imputación múltiple permiten modelar incertidumbre, aunque son poco usadas.

IDB Inter-American Development Bank

Servicio Nacional de Migraciones
Ministerio del Interior
Gobierno de Chile

ESTADÍSTICAS MIGRATORIAS