The Quiet Joining-Up

How governments link their citizens' records with Splink — and why citizens weren't really told

A public-interest report built entirely from publicly available government documents and peer-reviewed literature. No non-public systems were accessed. Compiled 2026-06-02. Confidence levels and sourcing are stated for every claim; what is unverified is flagged as such.

In one paragraph

Multiple national governments are using **Splink** — a free, open-source probabilistic record-linkage library built by the **UK Ministry of Justice** — to stitch previously-separate citizen datasets that share *no common identifier* into single, person-level linked records. This is not a leak, a hack, or a secret program: it is lawful, documented, and in several cases technically excellent. The public-interest question this report raises is narrower and more uncomfortable: **the people being linked were, as a rule, never individually told, never asked, and in most cases have no way to opt out** — because the legal basis is statutory "public task," not consent. The capability is also quietly migrating from *statistics* (counting populations) toward *operations* (identifying specific individuals in real time). That shift is the story.

1. What Splink is (and what it is not)

- Open-source library developed by the **MoJ data linking team**, funded by **ADR UK** under the **"Data First"** programme. Implements the **Fellegi–Sunter** model (the industry-standard statistical method for record linkage), estimated via Expectation-Maximisation, building on FastLink's R implementation. *(High confidence — GOV.UK; peer-reviewed IJPDS art. 1794, authored by the developers; official Splink docs; GitHub.)*
- Designed for **scale**: built to link **~100 million records**; MoJ has linked ~15 million in under an hour. Backend-agnostic (DuckDB / Spark / AWS Athena). *(High.)*
- **The maths is sound.** This report does **not** allege the algorithm is inaccurate or biased by design. Fellegi–Sunter is mainstream and the published accuracy is high. The concern is *governance of what the accurate output enables*, not a technical flaw.
- **This report is strictly about data-linkage governance.** It is unrelated to any software security issue.

2. Confirmed deployments (primary-sourced, adversarially verified)

🇬🇧 *UK — Ministry of Justice · "Data First" (High confidence)*

Splink links person records across **seven justice-system domains** that lack a consistent shared ID — Magistrates'/LIBRA, Crown/Common Platform, family courts, civil courts, prisons (NOMIS), probation (DELIUS), and offender-assessment systems — to track individuals' journeys through the system and identify repeat users. It deduplicates within each domain, then runs a further linkage to create a **whole-justice-system linked identifier (Cross Justice System / XJS table)**. Documented record counts: Common Platform 2.9m, LIBRA 19.6m, NOMIS 2.2m, DELIUS 2.4m. Scope: England & Wales. Researcher-facing data is de-identified.

> Sources: GOV.UK Algorithmic Transparency Records (*moj-data-first-splink*, *moj-splink-master-record*); ADR UK Data First flagship dataset; IJPDS 1794.

The function-creep flag: MoJ runs Splink in **batch** (weekly statistical refresh) *and* in **real time — "in courts, to find probation records associated with individuals coming to court."** It is piloting **"Core Person Record,"** a real-time system to assign a **unique cross-justice person identifier** across prisons, probation and criminal courts as records are created/updated. That is a move from *statistics* to *operational identification of named individuals*. (Core Person Record is described as a **pilot**, not yet operational — stated honestly.) *(High.)*

🇬🇧 *UK — Office for National Statistics (High confidence)*

ONS adopted Splink cross-government and is named in the UN Statistics Division catalogue. In the **2021 Census** linkage, ONS linked the Census to **DWP administrative data**: **96.7%** of census IDs (56,673,658 records) linked to a DWP master key or **encrypted National Insurance number**, the encrypted NINo also bridging to **HMRC PAYE** data; precision 99.87%, recall 99.86%. *Nuance, stated plainly:* within this specific exercise Splink was applied to the **non-linked records during false-negative analysis**, not as the primary engine. ONS also uses Splink for its **Business Index** (former IDBR) and **Demographic Index**.

> Source: ONS methodology, "2021 Census linkage to DWP master key and encrypted NINo" (Dec 2024); IJPDS 1794.

🇬🇧 *UK — NHS England (High confidence, but in-progress)*

NHS England generates a person-level identifier (**Person_ID**) via the **Master Person Service** (the live operational spine matching demographics to the NHS number) and is **"currently working on implementing a probabilistic linkage model using Splink"** as the engine for a new linkage service. *Stated honestly:* the Splink work is **in progress**, distinct from the already-operational Person_ID/MPS mechanism.

> Source: NHS England data science "Data Linkage Hub"; IJPDS 3271; NHS Digital MPS handbook.

🇦🇺 *Australia — ABS Person Linkage Spine (High confidence)*

"For the first time in 2025, the Spine was constructed using Splink." It links individuals across **three federal agencies**: **Medicare** (Services Australia), **Centrelink/DOMINO** (Dept. of Social Services), and **Personal Income Tax** (ATO) — health, welfare, and tax — covering the **ever-resident population Jan 2006–June 2025**. *(High.) Caveat: "ever-resident," slightly broader than "citizens."*

> Source: ABS, "Person Linkage Spine."

🇦🇺 *Australia — National Linkage Spine / National Disability Data Asset (High confidence)*

Built jointly by **ABS and AIHW** (co-technical leads) for the **NDDA**, performing **cross-jurisdictional** linkage to assemble person-level (de-identified) profiles of **people with disability**, using deterministic, probabilistic and ML methods, **"to be shared under new legislation"** enabling cross-jurisdictional sharing.

> Source: IJPDS 2818 (ABS/AIHW-authored); ABS/AIHW NDDA pages; ONDC data-sharing register DSR-03110.

---

3. The consent gap — the heart of the story *(High confidence, carefully scoped)*

Across **every** deployment reviewed, the public-facing documents cite governance artefacts — **DPIAs / DPIA screenings** (MoJ), **statutory "public task"** processing, **new data-sharing legislation** (Australia) — but contain **no mention of individual citizen consent, individual notification, or an opt-out.** The MoJ record's only individual-facing line is that data is *"routinely collected… for the purposes of managing court cases and offenders."* The ABS spine page offers only a generic "Five Safes / privacy and confidentiality" gesture.

This is the crux — and it must be stated fairly. The absence is **partly by design**: administrative-data and national-statistics processing legitimately rests on a **statutory/public-task** legal basis (e.g. the UK Digital Economy Act 2017; Australia's Census and Statistics Act / ABS Act), which **does not require consent**. So the finding is **not** "this is illegal" or "there is no legal basis." The finding is the genuine public-interest point: **the lawful basis is one most citizens have never heard of, the linkage is invisible to the individual, and there is no personal right to object** — even as the same machinery edges toward real-time operational use.

Honesty marker: these are **scoped absence-of-content findings** about the specific documents reviewed, **not** proof that no notification exists anywhere. ABS and MoJ both hold statutory authority documented on other pages.

4. The other side — why this exists, and the real safeguards *(for balance; a one-sided version is dismissible)*

- **Genuine public benefit:** accurate censuses, health research, reducing duplicate hospital records, understanding reoffending, building the evidence base for disability policy. These are not pretextual.
- **The Five Safes framework** (safe people/projects/settings/data/outputs) governs access. *(GOV.UK data ethics.)*
- **Accredited/Approved Researcher schemes** and the **ONS Secure Research Service** mean linked microdata is accessed in controlled environments by vetted researchers, with **de-identification/pseudonymisation** — researchers generally see linked records **without** direct identifiers. *(ONS; UK Statistics Authority; ADR UK ethics.)*
- **Oversight bodies** exist: the Office for Statistics Regulation, the Research Accreditation Panel, the ICO, and (Australia) the OAIC / National Data Commissioner.

The fair framing: the safeguards are real but **largely internal and researcher-facing**; they protect against *misuse of access*, not against the *prior question* of whether population-scale linkage should happen without the individual's knowledge — and they were built for the **statistics** use case, not the emerging **operational** one.

5. Explicitly UNVERIFIED — do not publish as fact

- **Credit scoring / credit bureaus: NOT substantiated.** No primary or secondary source found. **Excluded.** *(If anyone repeats this, it's the line that discredits the rest.)*
- **Asserted but not corroborated in this pass** (worth chasing, not yet citable here): UK Health Security Agency HIV-testing linkage; MoD Veterans Card / service-leavers↔Census; **US Defense Health Agency** (200M+ records); **Chile** migrant-immunisation linkage; **Statistics Canada / Stats NZ / Eurostat** adopters. The developers' phrase *"ONS and other collaborators"* implies more adopters — but "implies" is not "confirmed."

6. Open questions a journalist or regulator should put on the record

1. What is the **full** cross-government / international list of agencies running Splink on citizen data?
2. For the **pilot/operational** systems (MoJ Core Person Record; NHS England's Splink model; ABS NLS under "new legislation"), what is the timeline, and does operational use trigger **new notification or transparency duties**?
3. Is any individual **ever told** their records were probabilistically linked across health/justice/tax/benefits — and is there **any** opt-out, or is it public-task only with no right to object?
4. Which **independent** body (ICO, OSR, OAIC, National Data Commissioner) has actually audited these specific deployments and their **re-identification risk** — as opposed to the agencies' own DPIAs?

7. Sources (primary unless noted)

- GOV.UK — Splink: MoJ's open-source library for probabilistic record linkage at scale
- GOV.UK Algorithmic Transparency Records — *moj-data-first-splink*; *moj-splink-master-record*
- ADR UK — Data First (Cross Justice System, England & Wales); ADR UK ethics & responsibility
- IJPDS art. 1794 (Splink, developer-authored, peer-reviewed); art. 2818 (ABS/AIHW NLS/NDDA); art. 3271 (NHS probabilistic linkage)
- ONS — 2021 Census linkage to DWP master key and encrypted NINo (Dec 2024); Secure Research Service; Approved Researcher Scheme; public attitudes to data (June 2023)
- NHS England — Data Science "Data Linkage Hub"; NHS Digital Master Person Service
- ABS — Person Linkage Spine; National Disability Data Asset
- Splink official docs (Fellegi-Sunter) and GitHub (moj-analytical-services/splink)
- Office for Statistics Regulation — "Data sharing and linkage for the public good"
- GOV.UK — The Five Safes framework
- Georgina Sturge, *"The silent creep of data linkage"* (commentary/blog — framing reference, flagged as non-primary)

Verification: 6 search angles → 22 sources fetched → 102 candidate claims → 25 adversarially verified (2-of-3 refutation vote required to kill a claim) → 25 confirmed, 0 killed → synthesized to 9 findings.