## Informe comparativo Predio — Data Lake vs ColSmart (predios y atributos)

1. **Base de comparación (registros de predio)**

* **Data Lake: 518.088**
* **ColSmart: 518.588 → +500 (*“500 nuevo creado enviados por Carlos Fenandez”*).**

**Nota: En ColSmart hay 1 registro de numero\_predial más que el total de PREDIO (518.589 vs 518.588). Conviene 500 predio nuevo generados**

**2) Lectura ejecutiva**

* **Aumento del universo: ColSmart incorpora 500 predios adicionales frente a Data Lake**
* **Identificadores y claves:**
  + **numero\_predial alcanza cobertura total; (518.589 vs 518.588).**
  + **codigo\_orip mejora levemente (+501) tras la homologación desde matrícula.**
  + **matricula\_inmobiliaria disminuye 26,8% en ColSmart. Esto suele reflejar depuración/homologación (eliminación de duplicados, normalización o filtrado por validez). Requiere revisión dirigida.**
* **Atributos técnicos:**
  + **area\_catastral\_terreno no se migró (pérdida de 486.704 valores que en DL representaban 93,9% de cobertura). Es el mayor faltante.**
  + **tipo queda completo en ColSmart.**
  + **condicion\_predio aparece con 22.455 valores (4,33%), derivados de homologación desde numero\_predial.**
  + **destinacion\_economica mantiene cobertura muy alta (≈99,7% en ColSmart).**
  + **area\_registral\_m2 sin variación (≈4,2% de cobertura en ambos).**
* **Homologación del atributo “destino” se utilizó Corín generado por IA para homologar los destinos rurales** *(subconjunto con ese atributo)*:
  + **Destino homologado:** 516.853 (99,7%).
  + **Sin destino homologado:** 1.736 (0,3%).
  + **Base considerada:** 518.589.

**Lectura global:** **destino homologado (99,7% en su universo)**.

## Unidades de construcción y calificacion

**Corte analizado:** 591.619 unidades de construcción (según totales observados en las tablas suministradas).

**1) Resumen ejecutivo**

* **Cobertura y conciliación general (geo–alfanumérico):**
  + **Conciliadas (Geo + Alfa):** 364.322 (≈62%).
  + **Registros alfanuméricos sin geometría:** 86.091 (≈15%).
  + **Registros geográficos sin vínculo alfanumérico:** 141.206 (≈24%).
  + **Total:** 591.619 (100%).
* **Estado por criterios de calidad** (calificación tipológica, vínculo a predio y geometría válida):
  + **Listas para producción (✓ calificación, ✓ predio, ✓ geometría):** 259.250 (44%).
  + **Falta geometría válida (✓ calificación, ✓ predio, ✗ geometría):** 105.072 (18%).
  + **Sin tipología (✗ calificación, ✓ predio, ✓ geometría):** 185.704 (31%).
  + **Sólo geometría válida (✗ calificación, ✗ predio, ✓ geometría):** 41.593 (7%).
  + **Total:** 591.619 (100%).
* **Cobertura de “edad” de la construcción:**
  + **Con edad:** 551.968 (93%).
  + **Sin edad:** 39.651 (7%).
  + **Total:** 591.619.
* **Homologación del atributo “destino”** *(subconjunto con ese atributo)*:
  + **Destino homologado:** 516.853 (99,7%).
  + **Sin destino homologado:** 1.736 (0,3%).
  + **Base considerada:** 518.589.

**Lectura global:** El 44% ya está listo para producción. El principal pendiente es **calificar tipología/uso** (31%), seguido por **levantar/corregir geometría** (18%) y **asociar predio a huérfanas** (7%). La cobertura de atributos clave es alta: **edad (93%)** y **destino homologado (99,7% en su universo)**.

## Terrenos (migración y calidad de datos)

**Informe gerencial — Terrenos (actualizado)**

**Universo analizado:** **414.068** terrenos.

| **Categoría** | **Cantidad** | **% sobre el total** |
| --- | --- | --- |
| **Migrados correctamente** | **409.002** | **98,78 %** |
| **Geográfica – Terreno sin predio** | **5.054** | **1,22 %** |
| **Geográfica – Terreno NPM (longitud < 30 m)** | **12** | **0,003 %** |
| **Total** | **414.068** | **100 %** |

**Lectura ejecutiva:**

* **Listos:** 98,78 % de los terrenos migraron correctamente.
* **Brecha total a corregir:** **1,22 %** (5.066 terrenos), concentrada en **“sin predio”**; los **NPM < 30 m** son casos **puntuales** (12).

**Prioridades y acciones**

1. **Terreno sin predio (5.054 | 1,22 %)**
   * **Riesgo:** quiebra de integridad referencial y trazabilidad.
   * **Acción:** vincular mediante cruce alfanumérico y, si aplica, validación espacial (contain/overlap) contra la capa de predios; activar **FK obligatoria**.
2. **NPM (longitud < 30 m) (12 | 0,003 %)**
   * **Riesgo:** linderos anómalos / geometrías degeneradas.
   * **Acción:** revisión topológica y corrección manual o *snap* a cartografía base.
3. **Migrados correctamente (409.002 | 98,78 %)**
   * **Acción:** **monitoreo**; no requieren retrabajo.

## Interesados (homologación y enriquecimiento)

**Corte y alcance:** resultados del proceso de homologación de nombres (372.708 registros cruzados con resgistraduria) y su aplicación sobre la **tabla de interesados** (682.518 personas naturales).

**1) Resumen ejecutivo**

* **Automatización del proceso de homologación:** **95,35 %** de los 372.708 nombres fueron homologados automáticamente (Reglas léxicas + IA). Sólo **4,65 %** quedó pendiente de revisión manual.
* **Eficacia de IA en casos difíciles:** del universo con **lev > 2** (25.340), la IA homologó **8.023** → **31,66 %** de resolución automática en esa franja.
* **Impacto en la base maestra:** en la **tabla de interesados**, **620.000** de **682.518** registros se actualizaron con los algoritmos → **Cobertura de actualización: 90,84 %**.
* **Backlog actual:** **62.518** interesados (9,16 %) sin actualizar.
* **Ahorro operativo (referencia):** si una validación manual tarda **1 min/registro**, la automatización evitó **≈ 5.923 horas** de revisión (≈ 740 jornadas de 8 h). Con **30 s/registro**, el ahorro es **≈ 2.962 horas** (≈ 370 jornadas).

**2) KPIs gerenciales (fotografía del corte)**

| **KPI** | **Valor** |
| --- | --- |
| Universo de nombres a homologar registraduria | **372.708** |
| **Automatizados** (lev=0,1,2 + IA) | **355.391 (95,35 %)** |
| **Pendientes** (lev > 2 sin IA) | **17.317 (4,65 %)** |
| Eficacia IA en *casos difíciles* (lev > 2) | **31,66 %** (8.023/25.340) |
| Interesados PN en base maestra | **682.518** |
| **Interesados actualizados** | **620.000 (90,84 %)** |
| **Backlog de interesados** | **62.518 (9,16 %)** |
| Ahorro operativo (1 min/registro) | **≈ 5.923 h** (≈ 740 jornadas) |
| Ahorro operativo (30 s/registro) | **≈ 2.962 h** (≈ 370 jornadas) |

**Lectura:** El proceso está **maduro** (95 % de automatización) y ya **impactó** al 91 % de la base de interesados. El foco gerencial debe estar en **cerrar el 9 % restante** y **maximizar la eficacia IA** en pendientes de alta complejidad.

**3) Resultados operativos (detalle técnico)**

**Distribución por distancia de Levenshtein (372.708):**

| **Rango** | **Registros** | **%** | **¿Se homologó automáticamente?** |
| --- | --- | --- | --- |
| **lev = 0** | 326.431 | 87,58 % | Sí |
| **lev = 1** | 17.338 | 4,65 % | Sí (validado al 100 %) |
| **lev = 2** | 3.599 | 0,97 % | Sí (validado al 100 %) |
| **lev > 2** **con IA** | 8.023 | 2,15 % | Sí |
| **lev > 2** **pendiente** | 17.317 | 4,65 % | No (revisión) |
| **Total** | **372.708** | **100 %** |  |

**Conciliación con la tabla de interesados (PN):** 682.518 totales; 620.000 actualizados → **90,84 %** de cobertura; **62.518** pendientes (**9,16 %**).

**Alcance:** actualización del campo colsmart\_gender en la tabla de interesados, con trazabilidad del **método de asignación** (colsmart\_gender\_type).  
**Totales PN (base maestra, corte previo):** 682 518 personas naturales.

**1) Resumen ejecutivo**

* **Cobertura de género (personas naturales):**
  + **Registros procesados con alguna técnica:** **479 569** (**70,26 %** del total PN).
  + Dentro de los procesados, **con género asignado (F/M):** **472 268** (**98,48 %** de los procesados = **69,19 %** de la base PN).
  + **Pendientes dentro de los procesados** (*NO\_CLASIFICABLE + INDETERMINADO*): **7 301** (**1,52 %** de los procesados = **1,07 %** de la base PN).
  + **Sin procesar aún:** **202 949** (**29,74 %** del total PN).
* **Balance por género (sobre 479 569 procesados):**
  + **Femenino:** **235 406** (**49,09 %**).
  + **Masculino:** **236 862** (**49,39 %**).
  + **No clasificable:** **7 259** (**1,51 %**).
  + **Indeterminado:** **42** (**0,01 %**).
  + **Índice de equilibrio F/M:** 1,006 → **práctico equilibrio** (diferencia ~0,30 pp).
* **Contribución por método (sobre 479 569 procesados):**
  + **RANGO\_CEDULA:** **434 385** (**90,58 %**).
  + **COMPARACION\_EXACTA:** **24 463** (**5,10 %**).
  + **MODELO\_IA:** **20 721** (**4,32 %**).
  + **De los F/M asignados** (472 268): **92,0 %** proviene de **RANGO\_CEDULA**, **5,18 %** de **COMPARACION\_EXACTA** y **2,84 %** de **MODELO\_IA**.

**Lectura:** La cobertura ya permite análisis agregados (≈70 % de la base PN con género). El resultado muestra **paridad** F/M. Los pendientes se concentran en **NO\_CLASIFICABLE/INDETERMINADO** del **MODELO\_IA** y en el **29,7 %** aún no procesado.