DOCUMENTO 51 · DATOS Y CONOCIMIENTO

Datos y conocimiento para IA

Gobierno del dato aplicado a la IA, preparación por fase, linaje, datos personales, fuentes de conocimiento y propiedad intelectual

Fernando García Varela Versión 0.1 (borrador de trabajo) 16-09-2026 Borrador para revisión

2	8	19	12
esferas habilitadoras: datos y conocimiento	fases con requisitos de datos	indicadores con fórmula	errores frecuentes identificados

EN ESTE DOCUMENTO

01 Objeto y alcance

02 Esferas 05 y 06

03 Principios

04 Gobierno del dato aplicado a la IA

05 Preparación de datos por fase

06 Sistemas de alto riesgo: art. 10 del Reglamento Europeo de IA

07 Linaje de datos y modelos (P16)

08 Datos sintéticos

09 Datos personales en entrenamiento, instrucciones y resultados

10 Conocimiento

11 Propiedad intelectual y derechos de terceros

12 Indicadores

13 Errores frecuentes

14 Riesgos tipo de datos y conocimiento

15 Herramientas y plantillas asociadas

16 Documentos relacionados

17 Control de versiones

Aviso legal y exención de responsabilidad. SEVEN-G es un marco metodológico de referencia que se ofrece «tal cual» y con fines exclusivamente informativos. No constituye asesoramiento jurídico, regulatorio, financiero ni profesional, ni garantiza el cumplimiento de ninguna norma. Las referencias a regulación general (como el Reglamento Europeo de IA, el RGPD, DORA o NIS2), a normas técnicas y a regulación específica de cada sector o jurisdicción pueden ser incompletas, no aplicar a un caso concreto o quedar desactualizadas por cambios normativos, interpretaciones o criterios de las autoridades posteriores a su fecha de consulta. **Cada organización que use SEVEN-G es la única responsable de identificar la normativa que le aplica, verificar su vigencia y certificar su propio cumplimiento regulatorio**, con el asesoramiento cualificado que corresponda. Esta metodología es una ayuda genérica y gratuita, compartida con la comunidad para que nadie tenga que empezar desde cero; cada persona u organización puede y debe adaptarla a su propio uso. No debe entenderse que sus partes jurídicamente sensibles hayan sido revisadas por una asesoría jurídica: esas revisiones, para cada empresa o sector, son responsabilidad última de la empresa, el consultor o la organización que la use. Aunque se procura mantenerla al día, alguna norma puede haber cambiado sin que se recoja aquí. En la máxima medida permitida por la ley, el autor no asume responsabilidad alguna por los efectos de su aplicación en ninguna organización ni por su aplicabilidad completa. La metodología no otorga certificación de ningún tipo. El autor no asume responsabilidad alguna por el uso que se haga de este contenido ni por las decisiones que se adopten con él. Los datos, cifras, compañías y casos de los ejemplos son ficticios o ilustrativos.

LO ESENCIAL **Condicional: depende de un disparador**

Disparador: el caso de uso usa datos propios para entrenar, ajustar o alimentar un sistema, usa fuentes de conocimiento o trata datos personales. El linaje de datos y modelos (P16) es evidencia de todo caso de uso en la fase 4.

Matriz de obligatoriedad y lectura por capas (documento 94)

01

Objeto y alcance

Este documento desarrolla las esferas habilitadoras **05 · Datos** y **06 · Conocimiento** del mapa de impacto. Fija lo que una organización que aplica SEVEN-G debe hacer para que los datos y el conocimiento que usa la IA sean **conocidos, de calidad suficiente, utilizables legalmente, trazables y protegidos**.

Cubre:

- El gobierno del dato **aplicado a la IA**: propiedad, calidad, catálogo, base legal, minimización y retención.
- Los requisitos de datos en **cada fase del ciclo de vida**, con especial atención a la evaluación de disponibilidad y calidad en la fase 3.
- Los requisitos de gobernanza de datos del **art. 10 del Reglamento Europeo de IA** para sistemas de alto riesgo, en síntesis.
- El **linaje de datos y modelos** (P16).
- Los **datos sintéticos**.
- Los **datos personales** en entrenamiento, en instrucciones y en resultados.
- El **conocimiento** como fuente de la IA generativa: fuentes, curación, vigencia, permisos y conocimiento crítico que depende de pocas personas.
- La **propiedad intelectual** y los derechos de terceros.

No sustituye la política de gobierno del dato ni la de protección de datos de la compañía: las aplica a la IA y señala dónde la IA exige más. La seguridad de los datos frente a ataques se trata en el documento 35; los contratos con proveedores, en el documento 36.

Este documento no constituye asesoramiento jurídico.

02

Esferas 05 y 06

2.1 Preguntas de referencia

Esfera	Pregunta para el consejo
05 · Datos	¿Sabemos qué datos tenemos, quién responde de su calidad y si podemos usarlos legalmente para IA?

Esfera	Pregunta para el consejo
06 · Conocimiento	¿Cuánto conocimiento crítico depende de pocas personas y no está disponible para la organización?

Las dos esferas son **habilitadoras**: rara vez generan valor por sí solas, pero sin ellas los casos de uso de las esferas 01 a 04 fracasan o no se pueden defender ante un auditor o un regulador.

2.2 Niveles de ambición

Los ejemplos son **ilustrativos**.

Nivel	05 · Datos	06 · Conocimiento
Optimizar	Limpieza, normalización y deduplicación asistidas; catalogación con metadatos generados automáticamente y revisados; monitorización continua de la calidad.	Búsqueda semántica sobre documentación interna; bases de preguntas frecuentes mantenidas con ayuda de IA; resúmenes de reuniones que conservan decisiones.
Aumentar	Datos sintéticos donde los reales son escasos o sensibles; enriquecimiento con fuentes externas gobernadas; consulta de datos en lenguaje natural con controles de acceso.	Captura de conocimiento tácito de expertos; asistentes de conocimiento por dominio (jurídico, técnico, regulatorio, producto); detección de patrones de éxito y fracaso en la organización.
Transformar	Los datos se convierten en producto o servicio; arquitectura de datos federada con gobierno automatizado; representación actualizada de la operación completa.	El conocimiento organizativo se usa como capacidad activa y diferencial; la compañía conserva y reutiliza lo que ha hecho, decidido y aprendido.

En la esfera 05, el nivel **Transformar** exige especial atención a la base legal, a los derechos de terceros y a la clasificación regulatoria, porque cambia la finalidad para la que se recogieron los datos.

2.3 Qué se evalúa

En la esfera 05 se evalúan **propiedad** (DAT-01), **catálogo** (DAT-02), **legitimidad** (DAT-03), **calidad** (DAT-04, DAT-05, DAT-09), **trazabilidad** (DAT-06), **preparación** (DAT-07, DAT-10) y **retención** (DAT-08). En la esfera 06, **fuentes** (CNC-01, CNC-02, CNC-05), **calidad de uso** (CNC-03, CNC-04, CNC-08) y **dependencia de pocas personas** (CNC-06, CNC-07). Los indicadores se definen en la sección 12.

Estas dimensiones alimentan **D3 Datos y conocimiento** en el modelo de madurez (documento 11).

03

Principios

#	Principio	Qué implica
1	**Ningún dato sin propietario**	Todo conjunto de datos o fuente de conocimiento que usa un sistema de IA tiene un propietario de negocio identificado.
2	**La calidad se define para el uso**	Un dato suficiente para un informe puede ser insuficiente para entrenar un modelo. Los umbrales se fijan por caso.
3	**Usar datos para IA es una finalidad**	La base legal y la compatibilidad de finalidad se analizan antes de usar los datos, no después.
4	**Lo mínimo necesario**	Se usan solo los datos y los atributos necesarios, durante el tiempo necesario.
5	**Todo es trazable**	Se puede reconstruir qué datos, qué versión de modelo y qué fuentes produjeron un resultado.
6	**La IA no amplía permisos**	Un sistema de IA no da acceso a información que la persona que lo usa no podría ver por otra vía.
7	**El conocimiento caduca**	Toda fuente de conocimiento tiene fecha de revisión y un responsable de su vigencia.
8	**Los derechos de terceros se respetan**	No se usan datos ni contenidos sin título suficiente para el uso concreto.

04

Gobierno del dato aplicado a la IA

4.1 Roles

Rol	Responsabilidad en IA
Propietario del dato (negocio)	Autoriza el uso para un caso de uso concreto (en G3), define la calidad aceptable y responde de la finalidad.
Custodio del dato (tecnología)	Aplica controles de acceso, calidad, retención y copia; apoya al responsable técnico.
Responsable técnico de IA	Responde de los datos y modelos del caso de uso, de su preparación y de su linaje.

Rol	Responsabilidad en IA
Delegado de protección de datos	Asesora y supervisa en tratamientos de datos personales; consultado en la evaluación de impacto (fases 3–4); protección de datos interviene en la firma multinivel de G5.
Responsable de contenido o de conocimiento	Calidad, vigencia y permisos de las fuentes de la IA generativa (documento 50, sección 4.2).
Responsable de riesgos de IA	Valora los riesgos RT-DAT y emite conformidad.

4.2 Ficha de conjunto de datos para IA

Todo conjunto de datos que se usa para entrenar, ajustar, validar, probar o alimentar en producción un sistema de IA **debe** estar catalogado con, al menos, los campos siguientes. Si la compañía ya dispone de un catálogo de datos, la ficha se implementa como extensión de ese catálogo.

Campo	Contenido
Identificador, nombre, propietario y custodio	Principio 1.
Descripción y origen	Qué contiene, de qué sistemas procede, cómo se generó; si es sintético (sección 8).
Periodo y volumen	Fechas cubiertas y número de registros.
Datos personales	Sí o no; categorías, categorías especiales, menores; revisión del delegado de protección de datos.
Base legal y finalidades autorizadas	Por finalidad, con análisis de compatibilidad si es nueva (sección 9).
Restricciones de uso y confidencialidad	Contractuales, de licencia, sectoriales, territoriales (sección 11); clasificación.
Calidad	Dimensiones medidas, resultado y fecha (sección 4.3).
Representatividad y sesgos conocidos	Poblaciones cubiertas e infrarrepresentadas; obligatorio en alto riesgo (sección 6).
Retención	Plazo y regla de eliminación o anonimización (sección 4.5).
Sistemas que lo usan	Casos de uso y versiones de modelo; enlace a P16 y T02.
Versión	Versión y fecha de congelación si se usó para entrenar.

4.3 Calidad

Dimensión	Pregunta	Ejemplo de regla
Exactitud	¿Refleja la realidad?	Muestra verificada contra la fuente primaria.
Completitud	¿Faltan valores o registros?	Proporción de nulos por atributo crítico.
Coherencia	¿Es consistente entre sistemas y en el tiempo?	Conciliación de totales entre origen y destino.
Actualidad	¿Está al día para el uso?	Antigüedad máxima del último registro.
Validez	¿Cumple formato y dominio?	Valores dentro de listas y rangos permitidos.
Unicidad	¿Hay duplicados?	Registros duplicados por clave de negocio.
Representatividad	¿Cubre la población y las situaciones en que se usará el sistema?	Distribución por segmentos frente a la población de uso.
Calidad de etiquetas	¿Son correctas las etiquetas usadas para aprender?	Acuerdo entre etiquetadores en una muestra.

Los **umbrales** se fijan en la fase 3 para cada conjunto crítico y se convierten en controles de monitorización en la fase 6 (documento 52, sección 5).

4.4 Base legal y minimización

- El uso de datos personales para desarrollar u operar un sistema de IA es un tratamiento con **finalidad propia**. Debe tener base legal (RGPD, art. 6) y, si los datos se recogieron para otra finalidad, un análisis de **compatibilidad** (art. 6.4) o una base nueva.
- Las **categorías especiales** de datos (art. 9) requieren una excepción aplicable y controles reforzados.
- La **minimización** (art. 5.1.c) se aplica a atributos, registros y periodo. En IA se justifica por escrito cada atributo personal usado como variable.
- La **protección de datos desde el diseño y por defecto** (art. 25) se refleja en P15, P16 y P18.
- El tratamiento se incorpora al **registro de actividades de tratamiento** (art. 30) y, cuando proceda, a una **evaluación de impacto** (art. 35) en P11.

4.5 Retención

Elemento	Regla
Datos de entrenamiento, validación y prueba	Se conservan congelados y versionados mientras el modelo esté en uso y durante el plazo necesario para reconstruir o auditar, dentro de los límites de la base legal. Si no pueden conservarse, se conserva la descripción suficiente para la trazabilidad.
Datos de entrada y resultados en producción	Según el plazo de conservación de registros (documento 52, sección 11) y la minimización.
Instrucciones y conversaciones con IA generativa	Plazo definido y comunicado; no se conservan indefinidamente por defecto.
Índices y representaciones vectoriales	Siguen el plazo de la fuente de la que proceden; si la fuente se elimina, se elimina su representación.
Al retirar un sistema	Se aplica el tratamiento de datos y modelos decidido en G7 (documento 52, sección 12).

05

Preparación de datos por fase

5.1 Requisitos por fase

Fase	Qué se hace con los datos	Evidencia	Pregunta de parada
0 · Contexto	Declarar restricciones de datos conocidas: datos protegidos, transferencias internacionales, secretos, licencias.	P02	¿Hay una restricción que impide el caso de uso?
1 · Descubrimiento	Identificar qué datos o conocimiento serían necesarios y si existen.	P06	¿Depende la oportunidad de datos que no existen?
2 · Hipótesis	Confirmar que la línea base se puede medir con datos reales.	P09	¿Se puede medir la línea base?
3 · Viabilidad	**Evaluación de disponibilidad y calidad con datos reales**; base legal; derechos de terceros; representatividad.	P10, P11	¿Hay datos suficientes, de calidad suficiente y utilizables legalmente?
4 · Diseño	Linaje, ficha de conjuntos, controles de calidad, minimización, retención, permisos de fuentes de conocimiento.	P15, P16, P18	¿El diseño trata los datos con controles suficientes?

Fase	Qué se hace con los datos	Evidencia	Pregunta de parada
5 · Entrega	Congelar y versionar conjuntos; pruebas de calidad, sesgo y fuga de datos; validar fuentes de conocimiento.	P21, P22	¿Los resultados se sostienen con los datos de producción?
6 · Operación	Monitorizar calidad, deriva y vigencia; gestionar incidentes de datos.	P25, P27	¿Se ha degradado la calidad o ha cambiado la base legal?
7 · Evolución o retirada	Revisar si los datos permiten escalar; tratar datos y modelos al retirar.	P30	¿Los datos soportan el alcance ampliado?

5.2 Fase 3: la evaluación que más paradas evita

SEVEN-G parte de la premisa de que la **falta de disponibilidad o de calidad de los datos** es la principal causa de fracaso de los casos de uso de IA que no llegan a producción o no sostienen su valor. El marco no apoya esta premisa en cifras de mercado, sino en su lógica: un problema de datos descubierto en la fase 5 cuesta la construcción entera. Por eso la evaluación se sitúa en la fase 3, antes de la inversión en diseño y construcción, y el indicador DAT-07 permite contrastar la premisa con los datos propios de cada compañía.

Regla. La evaluación de datos de la fase 3 **debe** hacerse **con datos reales** o con una muestra representativa extraída de los sistemas de origen, no con descripciones de lo que los datos deberían contener. Una evaluación basada solo en entrevistas o en diccionarios de datos no supera G3.

Bloque	Comprobaciones
Existencia y acceso	Los datos existen; se ha obtenido acceso autorizado a una muestra; se conoce el coste y el plazo de acceso completo.
Volumen e historia	Hay registros y periodo suficientes para el tipo de sistema; hay casos suficientes de las situaciones relevantes, incluidas las minoritarias.
Calidad medida	Resultados por dimensión (sección 4.3) sobre la muestra, frente a umbrales propuestos.
Etiquetas o verdad de referencia	Existe forma de saber si el sistema acierta: etiquetas históricas, revisión experta o resultado observable, y con qué retraso.
Representatividad	La muestra cubre la población y los contextos de uso; sesgos conocidos documentados.
Legitimidad	Base legal y finalidad; análisis de compatibilidad; categorías especiales; transferencias; licencias y derechos de terceros.
Disponibilidad en producción	Los datos usados para construir estarán disponibles en producción con la misma definición, frecuencia y latencia.

Bloque	Comprobaciones
Conocimiento (IA generativa)	Fuentes identificadas, con propietario, vigencia y permisos conocidos; proporción de contenido desactualizado.
Coste de preparación	Esfuerzo estimado de limpieza, integración, etiquetado y curación, incluido en el coste completo (documento 42).

Resultado. Cada bloque se califica como *suficiente*, *suficiente con acciones* (acción, coste, plazo y responsable) o *insuficiente*. Un bloque de legitimidad *insuficiente* impide continuar. Si otros bloques son insuficientes, el resultado esperado de G3 es **Iterar**, **Pivotar** o **Parar** con motivo codificado *Datos insuficientes* (documento 03, sección 3.3).

06

Sistemas de alto riesgo: art. 10 del Reglamento Europeo de IA

Esta sección resume, sin detalle jurídico, lo que el art. 10 del Reglamento Europeo de IA exige a los conjuntos de datos de entrenamiento, validación y prueba de los **sistemas de alto riesgo**, y dónde se refleja en SEVEN-G. El análisis obligación por obligación, las fechas de aplicación (anexo III desde el 2 de diciembre de 2027 y anexo I desde el 2 de agosto de 2028, tras el Reglamento (UE) 2026/1744) y las demás modificaciones se tratan en el documento 34 (§3.1 y §3.2).

Requisito en síntesis	Dónde se refleja en SEVEN-G
Prácticas de gobernanza y gestión de datos adecuadas a la finalidad: decisiones de diseño, origen y finalidad original de los datos, preparación (anotación, etiquetado, limpieza, actualización, enriquecimiento, agregación), hipótesis sobre lo que los datos miden y representan.	Ficha de conjunto de datos (4.2); P16.
Evaluación previa de la disponibilidad, cantidad e idoneidad de los conjuntos de datos necesarios.	Evaluación de fase 3 (5.2); P10.
Examen de posibles sesgos que puedan afectar a la salud, la seguridad o los derechos fundamentales o dar lugar a discriminación, y medidas para detectarlos, prevenirlos y mitigarlos.	P22 (pruebas de sesgo); documento 50, sección 8.4; P12.
Detección de lagunas o deficiencias en los datos y forma de subsanarlas.	Evaluación de fase 3; acciones registradas.
Conjuntos de datos pertinentes, suficientemente representativos y, en la mayor medida posible, sin errores y completos en vista de la finalidad prevista.	Calidad (4.3) con umbrales por caso.

Requisito en síntesis	Dónde se refleja en SEVEN-G
Consideración de las características del contexto geográfico, conductual, contextual o funcional de uso.	Bloque de representatividad y disponibilidad en producción (5.2).
Tratamiento excepcional de categorías especiales de datos personales para detectar y corregir sesgos, con condiciones y garantías estrictas.	Sección 9.4; decisión documentada con el delegado de protección de datos.

Cuando la compañía actúa como **responsable del despliegue** de un sistema de alto riesgo de un proveedor, debe asegurarse de que los **datos de entrada** que controla son pertinentes y suficientemente representativos para la finalidad prevista (art. 26.4), y exigir al proveedor la información necesaria (documento 36).

Este documento no constituye asesoramiento jurídico.

07

Linaje de datos y modelos (P16)

7.1 Qué es y para qué sirve

El linaje permite responder, para cualquier resultado de un sistema de IA: **qué datos, qué transformaciones, qué versión de modelo, qué configuración y qué fuentes de conocimiento lo produjeron**. Es la base para investigar incidentes, atender derechos de las personas, auditar, reproducir resultados, gestionar cambios y retirar sistemas.

7.2 Contenido mínimo

Elemento	Contenido	Lite	Enterprise
Fuentes	Sistemas de origen, conjuntos de datos (con ficha) y fuentes de conocimiento.	Sí	Sí
Transformaciones	Pasos de preparación, reglas, código versionado.	Resumen	Detalle con versión
Conjuntos de entrenamiento, validación y prueba	Versión congelada, fecha, criterio de partición.	Si hay entrenamiento propio	Sí

Elemento	Contenido	Lite	Enterprise
Modelo	Tipo, versión, parámetros de configuración, fecha de entrenamiento, resultados de validación.	Sí	Sí
Modelo de terceros	Proveedor, identificador exacto de versión, fecha de adopción, condiciones de uso de los datos.	Sí	Sí
Instrucciones de sistema y plantillas (IA generativa y agentes)	Versión de las instrucciones, herramientas disponibles, filtros.	Sí	Sí
Índices de recuperación	Fuentes indexadas, fecha de indexación, parámetros de fragmentación y recuperación.	Si hay recuperación	Sí
Consumidores	Procesos, sistemas y decisiones que usan los resultados.	Sí	Sí
Relación con registros	Qué identificadores permiten enlazar un resultado concreto con las versiones anteriores.	Recomendado	Sí

7.3 Reglas

1. P16 se elabora en la fase 4 y **debe** estar vigente para G5.
2. Todo cambio en fuentes, transformaciones, modelo, instrucciones o índices **actualiza** P16 y se registra en P27 (documento 52, sección 6).
3. El linaje se genera **automáticamente** siempre que la plataforma lo permita; la documentación manual se limita a lo que no puede capturarse.
4. Un sistema en producción sin linaje vigente es una **no conformidad** (menor si no afecta a decisiones; mayor si el sistema es Enterprise).

08

Datos sintéticos

8.1 Cuándo tienen sentido

Uso	Ejemplo ilustrativo	Valor
Completar casos escasos	Situaciones poco frecuentes que el modelo debe reconocer.	Mejora de robustez.
Pruebas y desarrollo	Entornos de prueba sin datos personales reales.	Menor exposición.
Pruebas de estrés y seguridad	Casos adversos para evaluar un asistente o un agente.	Mejor evaluación.
Compartir con terceros	Proveedores que desarrollan sin acceso a datos reales.	Menor riesgo contractual y de privacidad.

8.2 Riesgos y controles

Riesgo	Control
Falsa sensación de anonimato: los datos sintéticos generados a partir de datos personales pueden permitir reidentificar o inferir información de personas reales.	Evaluación de riesgo de reidentificación antes de tratarlos como no personales; mientras no se demuestre, se tratan como datos personales.
Pérdida de fidelidad: no reproducen relaciones reales relevantes.	Validación de utilidad: comparación de distribuciones y del rendimiento de modelos entrenados con datos reales frente a sintéticos.
Amplificación de sesgos del conjunto de origen.	Pruebas de sesgo sobre el conjunto sintético y sobre el modelo resultante.
Validación circular: validar un modelo con datos generados por el mismo proceso que lo entrenó.	La validación final **debe** hacerse con datos reales o con una muestra real independiente.
Contaminación: mezcla no identificada de datos reales y sintéticos.	Marcado obligatorio del origen sintético en la ficha y en el linaje.

8.3 Reglas

1. Todo conjunto sintético tiene **ficha propia** (4.2) que indica el conjunto de origen, el método de generación y su versión.

2. Ningún resultado de validación de G5 se basa **solo** en datos sintéticos.
3. Los datos sintéticos en uso tienen validación de utilidad y de privacidad registrada (DAT-11).

09

Datos personales en entrenamiento, instrucciones y resultados

9.1 En entrenamiento y ajuste

Cuestión	Qué exige SEVEN-G
Finalidad y base legal	Documentadas en la ficha y verificadas por el delegado de protección de datos antes de G3.
Anonimización frente a seudonimización	Solo los datos efectivamente anonimizados quedan fuera del RGPD . La seudonimización reduce riesgo, pero los datos siguen siendo personales. La anonimización se justifica con un análisis de riesgo de reidentificación.
Memorización	Los modelos entrenados con datos personales pueden reproducirlos en sus resultados. Se prueba la extracción de datos en fase 5 cuando el modelo se entrena o ajusta con datos personales. El Comité Europeo de Protección de Datos ha analizado en su Dictamen 28/2024 cuándo un modelo de IA puede considerarse anónimo; se recomienda su consulta.
Derechos de los interesados	Se documenta cómo se atienden acceso, rectificación, supresión y oposición cuando los datos se han usado para entrenar, y qué implica para el modelo (reentrenamiento, filtros de salida u otras medidas).
Transparencia	La información a los interesados (RGPD, arts. 13 y 14) cubre el uso para IA cuando sea una finalidad nueva.

9.2 En instrucciones y contexto

Las instrucciones que los empleados escriben en herramientas de IA generativa, los documentos que adjuntan y el contexto que se recupera automáticamente son **datos que salen del control directo** de la persona y, a menudo, de la compañía.

Riesgo	Control
Empleados que introducen datos personales o confidenciales en herramientas no autorizadas.	Política de uso aceptable con clasificación de información permitida por herramienta (documento 31); controles de prevención de fuga de datos; monitorización del uso (T21).

Riesgo	Control
Proveedor que conserva instrucciones o las usa para entrenar sus modelos.	Condiciones contractuales verificadas en P14: no uso para entrenamiento, plazo de conservación, ubicación, subencargados (documento 36).
Recuperación automática de datos personales no necesarios para la respuesta.	Minimización en el diseño de la recuperación; filtros de datos personales en entrada y salida.
Instrucciones de sistema que contienen información sensible.	Las instrucciones de sistema no contienen secretos ni datos personales; se revisan en P18.
Registros de conversaciones conservados sin plazo.	Plazo definido (4.5) y acceso restringido.

9.3 En resultados

Los sistemas de IA pueden **inferir** datos personales que nadie ha facilitado (por ejemplo, rasgos, preferencias o situaciones sensibles). Una inferencia sobre una persona identificada es un dato personal. Si la inferencia corresponde a una categoría especial, se aplican sus reglas. El diseño (P17) **debe** identificar qué inferencias produce el sistema y si son necesarias y lícitas.

9.4 Datos para detectar sesgos

Detectar sesgos puede requerir datos sobre características protegidas que la compañía normalmente no trata. Para sistemas de alto riesgo, el Reglamento Europeo de IA admite de forma excepcional el tratamiento de categorías especiales con esa finalidad y con garantías estrictas (art. 10.5), y el Reglamento (UE) 2026/1744 extiende esa posibilidad más allá del alto riesgo con criterio de estricta necesidad (documento 34 §3.2). En todo caso, la decisión se toma con el delegado de protección de datos y el análisis jurídico, se documenta en P11 y se aplican medidas como separación, seudonimización, acceso restringido y eliminación al terminar el análisis.

10

Conocimiento

10.1 Fuentes de conocimiento para IA generativa

Cuando un sistema de IA generativa responde apoyándose en documentos y contenidos de la compañía mediante recuperación (*RAG*, generación aumentada por recuperación), **la calidad de las respuestas depende más de las fuentes que del modelo**. SEVEN-G trata cada fuente como un activo gobernado.

Inventario de fuentes. Cada sistema mantiene la relación de fuentes que usa, con:

Campo	Contenido
Fuente	Repositorio, espacio, base de conocimiento o sistema.
Responsable de contenido	Persona o función que responde de la calidad y vigencia.
Alcance	Qué partes se indexan y cuáles se excluyen expresamente.
Tipo de contenido	Normativo interno, procedimiento, producto, técnico, jurídico, comercial, histórico.
Autoridad	Fuente oficial, de apoyo o informativa; qué prevalece en caso de contradicción.
Clasificación y permisos	Confidencialidad y modelo de permisos de origen.
Vigencia	Fecha de revisión obligatoria por tipo de contenido.
Frecuencia de actualización del índice	Continua, diaria, semanal.
Derechos	Si incluye contenido de terceros con licencia (sección 11).

10.2 Criterios de admisión

Una fuente **no debería** conectarse a un sistema de IA generativa si:

- No tiene responsable de contenido.
- No puede determinarse qué documentos están vigentes y cuáles no.
- Mezcla borradores, versiones antiguas y documentos oficiales sin distinguirlos.
- Su modelo de permisos no puede trasladarse a la recuperación (sección 10.4).
- Incluye datos personales o contenido de terceros cuyo uso no se ha analizado.

10.3 Curación y vigencia

Actividad	Qué se hace	Frecuencia
Depuración inicial	Eliminar duplicados, versiones superadas y borradores; marcar el documento vigente.	Antes de G5
Metadatos mínimos	Responsable, fecha de publicación, fecha de revisión, estado (vigente, sustituido, histórico), autoridad.	Al incorporar cada documento

Actividad	Qué se hace	Frecuencia
Revisión de vigencia	El responsable de contenido confirma, actualiza o retira los documentos que alcanzan su fecha de revisión.	Según tipo; orientativamente, 6 meses para procedimientos y 12 para contenido estable
Retirada del índice	Los documentos sustituidos o caducados salen del índice o se marcan como históricos y se excluyen de la recuperación por defecto.	En cada actualización
Detección de lagunas	Análisis de consultas sin documento pertinente recuperado (CNC-08) para crear o mejorar contenido.	Mensual
Detección de contradicciones	Revisión de documentos que dan respuestas incompatibles sobre el mismo tema.	Trimestral

Regla. Un sistema que cita un documento caducado en una respuesta sobre condiciones, precios, normas o procedimientos que afectan a clientes o empleados genera un **incidente** que se clasifica con la escala S1–S4 (documento 52, sección 7).

10.4 Permisos de acceso

La recuperación de contenidos es una **vía de acceso a la información** y debe respetar los mismos permisos que la fuente.

Riesgo	Control
Sobreexposición: el asistente muestra a una persona contenido que no podría abrir en la fuente.	Filtrado por permisos del usuario **en el momento de la recuperación**, no solo al indexar; pruebas específicas en fase 5.
Permisos heredados excesivos en la fuente, que la IA hace visibles.	Revisión de permisos de la fuente antes de conectarla; la IA hace fácil encontrar lo que antes estaba "oculto por desconocido".
Índice común para colectivos con permisos distintos.	Índices separados o filtrado verificado por colectivo.
Agentes con permisos propios que recuperan información para usuarios con menos permisos.	Los agentes actúan con los permisos del usuario al que sirven, salvo diseño justificado en P18 (documento 35).
Cambios de permisos en la fuente no trasladados al índice.	Sincronización de permisos con plazo máximo definido y monitorizado.

10.5 Evaluación de la calidad del conocimiento en uso

Aspecto	Qué se mide	Indicador
Pertinencia de la recuperación	Si los documentos recuperados son los adecuados para la consulta.	Evaluado en el conjunto de evaluación (documento 52, sección 4.2).
Fidelidad a las fuentes	Si las afirmaciones de la respuesta están soportadas por los documentos recuperados.	CNC-04
Uso de contenido caducado	Si se citan documentos fuera de vigencia.	CNC-03
Lagunas	Consultas sin documento pertinente.	CNC-08
Cita de fuentes	Si la respuesta permite al usuario comprobar la fuente.	Requisito de diseño en P17.

10.6 Conocimiento crítico dependiente de pocas personas

Paso	Actividad	Resultado
1 · Identificar dominios críticos	Procesos, productos, sistemas o relaciones cuyo conocimiento es imprescindible para operar o decidir.	Lista de dominios críticos (C1).
2 · Medir la dependencia	Para cada dominio: cuántas personas lo dominan, si está documentado y utilizable, y riesgo de salida (jubilación, rotación, movilidad).	Mapa de dependencia (CNC-06).
3 · Priorizar	Dominios con dos personas o menos, sin documentación utilizable y con riesgo de salida en el horizonte de la tesis.	Plan de captura.
4 · Capturar	Entrevistas estructuradas (pueden asistirse con IA), grabación de explicaciones, revisión de casos, documentación de decisiones y excepciones.	Contenido con responsable y vigencia.
5 · Validar	El experto y un segundo revisor confirman que el contenido es correcto y completo para su uso.	Contenido validado.
6 · Poner a disposición	Incorporación como fuente gobernada (sección 10.1).	CNC-07

Consideraciones sobre las personas. La captura de conocimiento tácito se hace con **participación voluntaria informada**, reconocimiento del experto y claridad sobre el uso posterior del contenido. No **debe** presentarse ni utilizarse como paso previo a prescindir de la persona sin que el efecto se haya

evaluado conforme al documento 50. Si las grabaciones o transcripciones contienen datos personales, se aplican la base legal, la información y la retención correspondientes.

11

Propiedad intelectual y derechos de terceros

Consulta realizada en septiembre de 2026. Materia en evolución normativa y jurisprudencial; debe verificarse con asesoría especializada.

11.1 Datos y contenidos de entrada

Situación	Cuestión	Control
Datos adquiridos o licenciados	La licencia puede no permitir el uso para entrenar o para alimentar sistemas de IA, ni el uso por proveedores.	Revisión de licencia antes de G3; restricciones en la ficha (4.2).
Contenido obtenido de fuentes públicas en internet	Que un contenido sea accesible no significa que su uso sea libre. En la Unión Europea, la minería de textos y datos tiene excepciones con condiciones, y los titulares pueden reservarse ese uso para finalidades distintas de la investigación científica (Directiva (UE) 2019/790 , arts. 3 y 4).	Análisis jurídico previo; respeto de las reservas de derechos expresadas de forma adecuada; registro de fuentes.
Contenido de clientes o socios	Su uso puede estar limitado por contrato y por confidencialidad.	Verificación contractual; exclusión por defecto.
Secretos empresariales propios	Introducirlos en servicios de terceros puede comprometer las medidas razonables de protección exigidas para que sigan siendo secretos (Directiva (UE) 2016/943 ; en España, Ley 1/2019 de Secretos Empresariales).	Clasificación, herramientas autorizadas y condiciones contractuales.
Software y modelos de código abierto	Las licencias pueden imponer condiciones de uso, atribución o restricciones de finalidad.	Registro de licencias en P15; revisión antes de G4.

11.2 Modelos de propósito general de terceros

El Reglamento Europeo de IA obliga a los **proveedores de modelos de IA de propósito general** a establecer una política para cumplir la normativa de derechos de autor, incluidas las reservas de derechos, y a publicar un resumen suficientemente detallado del contenido usado para el entrenamiento (art. 53.1.c y d). La compañía que usa esos modelos **debería** pedir y conservar esa información en la evaluación del proveedor (P14) y valorar las garantías contractuales frente a reclamaciones de terceros (documento 36).

11.3 Resultados generados

Cuestión	Qué exige SEVEN-G
Titularidad	La protección por derechos de autor de contenidos generados con IA es incierta y depende de la jurisdicción y del grado de aportación humana. Los contenidos que la compañía necesita proteger se generan con aportación humana documentada o se protegen por otras vías (contrato, secreto).
Reproducción de obras de terceros	Los resultados pueden reproducir contenido protegido. En usos externos (marketing, producto, publicaciones) se aplican revisión humana y, cuando proceda, herramientas de detección.
Condiciones del proveedor	Se verifica qué derechos concede el proveedor sobre los resultados y qué uso puede hacer él de ellos.
Transparencia	Se cumplen las obligaciones de marcado o información sobre contenido generado o manipulado que resulten aplicables (documento 34).

Este documento no constituye asesoramiento jurídico.

12

Indicadores

Los códigos son provisionales; su código definitivo `IND-` figura en la tabla de correspondencia del documento 41 (§21). Los indicadores de conocimiento usan el prefijo **CNC-** para no confundirse con la familia `IND-CON` (consejo) del documento 41. Los umbrales son **orientativos y a calibrar**. "Sin dato" no es cero.

12.1 Esfera 05 · Datos

Código	Indicador	Fórmula	Frecuencia	Fuente	Referencia orientativa
DAT-01	Cobertura de propiedad	Conjuntos de datos usados por sistemas en producción con propietario asignado ÷ conjuntos usados por sistemas en producción	Trimestral	Catálogo, T02	100 %
DAT-02	Cobertura de catálogo	Conjuntos usados por IA con ficha completa ÷ conjuntos usados por IA	Trimestral	Catálogo	100 % en Enterprise

Código	Indicador	Fórmula	Frecuencia	Fuente	Referencia orientativa
DAT-03	Cobertura de base legal	Conjuntos con datos personales usados por IA con base legal y finalidad verificadas ÷ conjuntos con datos personales usados por IA	Trimestral	Catálogo, registro de tratamientos	100 %
DAT-04	Calidad por dimensión	Registros que superan las reglas de la dimensión ÷ registros evaluados (por conjunto y dimensión)	Según frecuencia de carga	Monitorización de calidad	Umbral fijado por caso en fase 3
DAT-05	Conjuntos críticos en umbral	Conjuntos críticos que cumplen todos sus umbrales ÷ conjuntos críticos	Mensual	Monitorización de calidad	≥ 95 %
DAT-06	Cobertura de linaje	Sistemas en producción con P16 vigente ÷ sistemas en producción	Trimestral	T02, P16	100 %
DAT-07	Decisiones de G3 por datos	Decisiones de G3 con resultado Iterar, Pivotar o Parar y motivo *Datos insuficientes* ÷ decisiones de G3	Semestral	T01, T03	Lectura, no objetivo: muy baja puede indicar evaluación laxa; muy alta, mala selección de oportunidades
DAT-08	Cumplimiento de retención	Conjuntos con plazo de retención definido y aplicado ÷ conjuntos usados por IA	Semestral	Catálogo	100 %
DAT-09	Incidentes con causa en datos	Incidentes de IA cuya causa raíz es de datos ÷ incidentes de IA cerrados	Trimestral	T08	Tendencia decreciente
DAT-10	Tiempo de acceso a datos	Mediana de días entre la solicitud de acceso a datos de un caso de uso y su disponibilidad autorizada	Trimestral	Solicitudes de acceso	Tendencia decreciente
DAT-11	Datos sintéticos validados	Conjuntos sintéticos en uso con validación de utilidad y privacidad registrada ÷ conjuntos sintéticos en uso	Semestral	Catálogo	100 %

12.2 Esfera 06 · Conocimiento

Código	Indicador	Fórmula	Frecuencia	Fuente	Referencia orientativa
CNC-01	Fuentes con responsable	Fuentes de conocimiento conectadas a IA con responsable de contenido ÷ fuentes conectadas	Trimestral	Inventario de fuentes	100 %
CNC-02	Vigencia del contenido	Documentos indexados dentro de su fecha de revisión ÷ documentos indexados	Mensual	Metadatos de fuentes	≥ 95 %
CNC-03	Uso de contenido caducado	Respuestas muestreadas que citan documentos fuera de vigencia ÷ respuestas muestreadas con cita	Mensual	Evaluación continua	Próximo a 0 %
CNC-04	Fidelidad a las fuentes	Respuestas evaluadas cuyas afirmaciones están soportadas por las fuentes recuperadas ÷ respuestas evaluadas	Continua o mensual	Evaluación continua	Umbral por caso; se comparte con OPE-09
CNC-05	Exposición indebida	Hallazgos de acceso a contenido sin permiso detectados en pruebas o en producción, por periodo	Trimestral	Pruebas de permisos, T08	0
CNC-06	Dependencia de conocimiento crítico	Dominios críticos con dos personas o menos que los dominan y sin documentación utilizable ÷ dominios críticos	Anual (C1, C5)	Mapa de dependencia	Tendencia decreciente
CNC-07	Cobertura de captura	Dominios críticos dependientes con contenido capturado y validado ÷ dominios críticos dependientes	Semestral	Plan de captura	Según plan
CNC-08	Consultas sin fuente pertinente	Consultas sin documento pertinente recuperado ÷ consultas evaluadas	Mensual	Evaluación continua	Tendencia decreciente

13

Errores frecuentes

#	Error	Consecuencia	Cómo se evita
1	Evaluar los datos en fase 3 con descripciones, no con datos reales.	Se descubre en fase 5 que no sirven.	Sección 5.2.
2	Suponer que los datos de un sistema sirven para cualquier finalidad.	Tratamiento sin base legal.	Legitimidad bloqueante en G3.
3	Construir con datos que no estarán igual en producción.	Rendimiento real muy inferior al validado.	Bloque de disponibilidad en producción.
4	No poder saber si el sistema acierta.	Imposible validar y monitorizar.	Bloque de etiquetas.
5	Conectar un repositorio entero a un asistente.	Borradores, versiones antiguas, sobreexposición.	Secciones 10.2 y 10.3.
6	Filtrar permisos al indexar y no al recuperar.	Empleados ven lo que no deberían.	Sección 10.4.
7	Tratar como anónimos datos seudonimizados o sintéticos.	Incumplimiento del RGPD.	Secciones 8.2 y 9.1.
8	No versionar los conjuntos de entrenamiento.	Imposible reproducir o investigar.	Secciones 4.2 y 7.
9	No actualizar el linaje.	Linaje falso ante un incidente.	Sección 7.3.
Mapa de esferas y niveles de ambición (10)	Usar contenido de terceros por estar en internet.	Reclamaciones de propiedad intelectual.	Sección 11.1.
Modelo de madurez (11)	Omitir el coste de preparación, etiquetado y curación.	Neto esperado inflado en G3.	Sección 5.2; documento 42.
Índice de transformación (12)	Capturar conocimiento de expertos sin su participación informada.	Contenido pobre y pérdida de confianza.	Sección 10.6; documento 50.

14

Riesgos tipo de datos y conocimiento

Códigos y nombres del catálogo de riesgos tipo del documento 33; dos de estos riesgos figuran allí en otras categorías (LEG, ORG).

Código (documento 33)	Riesgo	Controles principales
RT-DAT-01	Calidad insuficiente: datos insuficientes o de calidad inadecuada para la finalidad.	Evaluación de fase 3 con datos reales; umbrales y monitorización.
RT-DAT-03	Falta de base legal o finalidad incompatible en el uso de datos personales.	Ficha, verificación del delegado de protección de datos, evaluación de impacto.
RT-DAT-02	Datos no representativos y sesgo que producen resultados discriminatorios.	Representatividad, pruebas de sesgo, sección 9.4.
RT-DAT-04	Linaje desconocido: pérdida de trazabilidad de datos y modelos.	P16 vigente y actualizado con cada cambio.
RT-DAT-05	Exposición de datos en el ciclo de IA: fuga de datos personales o confidenciales por instrucciones, recuperación o memorización.	Secciones 9.2 y 10.4; pruebas de extracción.
RT-DAT-06	Conocimiento obsoleto o contradictorio en las respuestas.	Curación y vigencia; CNC-02 y CNC-03.
RT-LEG-06	Infracción de propiedad intelectual u otros derechos de terceros por datos o contenidos usados.	Sección 11; evaluación del proveedor.
RT-ORG-02	Dependencia de personas clave: pérdida de conocimiento crítico por salida de personas.	Sección 10.6; CNC-06 y CNC-07.

15

Herramientas y plantillas asociadas

Este documento no crea una herramienta nueva. La ficha de conjunto de datos y el inventario de fuentes de conocimiento se implementan en el **catálogo de datos que ya use la compañía** o, si no existe, en la plantilla P64 y como atributos del sistema en T02.

Nombre	Uso en este documento
Inventario, *gates*, riesgos, clasificador regulatorio (T02 · T03 · T06 · T07)	Enlace de sistemas con datos y linaje; evaluación de G3; riesgos RT-DAT; obligaciones de alto riesgo.
Incidentes, proveedores, uso corporativo (T08 · T09 · T21)	Incidentes con causa en datos; condiciones de proveedores; datos en herramientas no autorizadas.
Declaración de contexto y restricciones (P02)	Restricciones de datos de la fase 0.
Evaluación de viabilidad (P10)	Evaluación de disponibilidad y calidad de datos (sección 5.2).
Clasificación regulatoria y evaluaciones de impacto (P11)	Base legal, evaluación de impacto en protección de datos, datos para detectar sesgos.
Evaluación de proveedor de IA (P14)	Uso de instrucciones y datos por el proveedor; política de derechos de autor de modelos de propósito general.
Registro de arquitectura (P15)	Licencias de componentes y modelos.
Linaje de datos y modelos (P16)	Plantilla principal de este documento; usa las fichas y el inventario de fuentes de P64.
Diseño de seguridad (P18)	Permisos de recuperación y de agentes.
Resultados de validación y del piloto (P22)	Pruebas de calidad, sesgo, fuga de datos y permisos.
Configuración de monitorización y alertas (P25)	Umbrales de calidad de datos y de vigencia.
Ficha de conjunto de datos e inventario de fuentes de conocimiento (P64)	Ficha de cada conjunto de datos e inventario de fuentes de conocimiento, si la compañía no tiene catálogo de datos.

16

Documentos relacionados

Documento	Relación
00, 01 y 03 · Fundamentos y registro	Esferas 05 y 06; actividades de datos por fase; motivo *Datos insuficientes*.
10, 11 y 20 · Esferas, madurez y manuales de fase	Esferas 05 y 06; dimensión D3; requisitos de la sección 5 en cada fase.
31 · Política corporativa y uso aceptable	Información que puede introducirse en cada herramienta.

Documento	Relación
33 · Metodología de riesgos de IA	Consolidación de los riesgos RT-DAT.
34 · Mapeo regulatorio	Art. 10 del Reglamento Europeo de IA , RGPD y propiedad intelectual en detalle.
35 · Seguridad de IA y agentes	Protección de datos frente a ataques; permisos de agentes.
36 · Terceros y proveedores de IA	Cláusulas de uso de datos, confidencialidad y propiedad intelectual.
41 · Catálogo de indicadores	Consolidación de los indicadores DAT y CNC (tabla de correspondencia de 41 §21).
42 · Costes de IA	Coste de preparación, etiquetado y curación (categoría *datos*).
50 · Personas y organización	Captura de conocimiento de expertos y datos de empleados.
52 · Manual de operación de IA	Monitorización de calidad y deriva, registros, cambios y retirada.

17

Control de versiones

Versión	Fecha	Cambios
0.1	16-09-2026	Primera versión. Desarrolla las esferas 05 y 06 con indicadores; el gobierno del dato aplicado a la IA (roles, ficha de conjunto de datos, calidad, base legal, minimización y retención); los requisitos de datos por fase con la evaluación de fase 3 con datos reales; la síntesis del art. 10 del Reglamento Europeo de IA ; el contenido de P16; datos sintéticos; datos personales en entrenamiento, instrucciones y resultados; fuentes de conocimiento para IA generativa, curación, vigencia, permisos y conocimiento crítico dependiente de pocas personas; propiedad intelectual; diecinueve indicadores con fórmula, doce errores frecuentes y ocho riesgos tipo. Ajustes de coherencia con 01 (separación de funciones en Lite, resultados de R6, criterio de agentes) y con 34 y 37; riesgos alineados con los códigos del documento 33. Indicadores de conocimiento renombrados de CON- a CNC- para evitar la colisión con la familia IND-CON (consejo) del documento 41.

Fuentes externas citadas

Enlaces a la fuente oficial o primaria, comprobados antes de publicar. Verifique siempre la versión vigente en la fecha de uso.

Directiva (UE) 2016/943 del Parlamento Europeo y del Consejo, de 8 de junio de 2016, relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) contra su obtención, utilización y revelación ilícitas · https://eur-lex.europa.eu/eli/dir/2016/943/oj/spa · Parlamento Europeo y Consejo de la UE · 08-06-2016

Directiva (UE) 2019/790 del Parlamento Europeo y del Consejo, de 17 de abril de 2019, sobre los derechos de autor y derechos afines en el mercado único digital · https://eur-lex.europa.eu/eli/dir/2019/790/oj/spa · Parlamento Europeo y Consejo de la UE · 17-04-2019

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos) · https://eur-lex.europa.eu/eli/reg/2016/679/oj/spa · Parlamento Europeo y Consejo de la UE · 27-04-2016

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial) · https://eur-lex.europa.eu/eli/reg/2024/1689/oj/spa · Parlamento Europeo y Consejo de la UE · 13-06-2024

Reglamento (UE) 2026/1744 del Parlamento Europeo y del Consejo, de 8 de julio de 2026, por el que se modifican los Reglamentos (UE) 2024/1689, (UE) 2018/1139 y (UE) 2023/1230 en lo que respecta a la simplificación de la aplicación de normas armonizadas en materia de inteligencia artificial (Ómnibus digital sobre IA) · https://eur-lex.europa.eu/eli/reg/2026/1744/oj/spa · Parlamento Europeo y Consejo de la UE · 08-07-2026

Comité Europeo de Protección de Datos (CEPD) · https://www.edpb.europa.eu/edpb_es · Comité Europeo de Protección de Datos (organismo de la UE creado por el art. 68 del RGPD)

Dictamen 28/2024 sobre determinados aspectos de la protección de datos relacionados con el tratamiento de datos personales en el contexto de los modelos de IA · https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_es · Comité Europeo de Protección de Datos · 17-12-2024