DOCUMENT 51 · DATA AND KNOWLEDGE

Data and knowledge for AI

Data governance applied to AI, preparation by phase, lineage, personal data, knowledge sources and intellectual property

Fernando García Varela Version 0.1 (working draft) 16-09-2026 Draft for review

2	8	19	12
enabling spheres: data and knowledge	phases with data requirements	indicators with a formula	common mistakes identified

IN THIS DOCUMENT

01 Purpose and scope

02 Spheres 05 and 06

03 Principles

04 Data governance applied to AI

05 Data preparation by phase

06 High-risk systems: Article 10 of the EU AI Act

07 Data and model lineage (P16)

08 Synthetic data

09 Personal data in training, prompts and outputs

10 Knowledge

11 Intellectual property and third-party rights

12 Indicators

13 Common mistakes

14 Typical data and knowledge risks

15 Associated tools and templates

16 Related documents

17 Version control

Legal notice and disclaimer. SEVEN-G is a reference methodological framework provided "as is" and for information purposes only. It does not constitute legal, regulatory, financial or professional advice, nor does it guarantee compliance with any law or standard. References to general regulation (such as the EU AI Act, the GDPR, DORA or NIS2), to technical standards and to regulation specific to each sector or jurisdiction may be incomplete, may not apply to a particular case or may become out of date as a result of regulatory changes, interpretations or supervisory positions after their consultation date. **Each organisation that uses SEVEN-G is solely responsible for identifying the regulation that applies to it, verifying that it is current and certifying its own regulatory compliance**, with the appropriate qualified advice. This methodology is a generic, free aid shared with the community so that nobody has to start from scratch; each person or organisation can and should adapt it to its own use. It must not be inferred that its legally sensitive parts have been reviewed by legal counsel: those reviews, for each company or sector, are the ultimate responsibility of the company, consultant or organisation that uses it. Although every effort is made to keep it up to date, some regulation may have changed without being reflected here. To the fullest extent permitted by law, the author accepts no responsibility whatsoever for the effects of its application in any organisation or for its full applicability. The methodology does not grant certification of any kind. The author accepts no liability for any use made of this content or for any decisions taken on the basis of it. The data, figures, companies and cases in the examples are fictitious or illustrative.

THE ESSENTIALS Conditional: depends on a trigger

Trigger: the use case uses the company's own data to train, fine-tune or feed a system, uses knowledge sources or processes personal data. Data and model lineage (P16) is evidence for every use case in phase 4.

Obligation matrix and layered reading (document 94)

01

Purpose and scope

This document develops the enabling spheres **05 · Data** and **06 · Knowledge** of the impact map. It sets out what an organisation applying SEVEN-G must do to ensure that the data and knowledge used by AI are **known, of sufficient quality, legally usable, traceable and protected**.

It covers:

- Data governance **applied to AI**: ownership, quality, catalogue, legal basis, minimisation and retention.
- The data requirements in **each phase of the lifecycle**, with particular attention to the availability and quality assessment in phase 3.
- The data governance requirements of **Article 10 of the EU AI Act** for high-risk systems, in summary.
- **Data and model lineage** (P16).
- **Synthetic data**.
- **Personal data** in training, in prompts and in outputs.
- **Knowledge** as a source for generative AI: sources, curation, validity, permissions and critical knowledge that depends on a few people.
- **Intellectual property** and third-party rights.

It **does not replace** the company's data governance policy or its data protection policy: it applies them to AI and indicates where AI demands more. The security of data against attacks is dealt with in document 35; contracts with suppliers, in document 36.

This document does not constitute legal advice.

02

Spheres 05 and 06

2.1 Reference questions

Sphere	Question for the board
05 · Data	Do we know what data we have, who is accountable for its quality and whether we can legally use it for AI?

Sphere	Question for the board
06 · Knowledge	How much critical knowledge depends on a few people and is not available to the organisation?

Both spheres are **enablers**: they rarely generate value on their own, but without them use cases in spheres 01 to 04 fail or cannot be defended before an auditor or a regulator.

2.2 Ambition levels

The examples are **illustrative**.

Level	05 · Data	06 · Knowledge
Optimise	Assisted cleansing, normalisation and deduplication; cataloguing with automatically generated and reviewed metadata; continuous quality monitoring.	Semantic search over internal documentation; FAQ bases maintained with the help of AI; meeting summaries that preserve decisions.
Augment	Synthetic data where real data are scarce or sensitive; enrichment with governed external sources; natural language data querying with access controls.	Capture of experts' tacit knowledge; knowledge assistants by domain (legal, technical, regulatory, product); detection of patterns of success and failure in the organisation.
Transform	Data become a product or service; federated data architecture with automated governance; up-to-date representation of the entire operation.	Organisational knowledge is used as an active and differentiating capability; the company retains and reuses what it has done, decided and learned.

In sphere 05, the **Transform** level requires particular attention to the legal basis, third-party rights and regulatory classification, because it changes the purpose for which the data were collected.

2.3 What is assessed

In sphere 05, the following are assessed: **ownership** (DAT-01), **catalogue** (DAT-02), **legitimacy** (DAT-03), **quality** (DAT-04, DAT-05, DAT-09), **traceability** (DAT-06), **readiness** (DAT-07, DAT-10) and **retention** (DAT-08). In sphere 06: **sources** (CNC-01, CNC-02, CNC-05), **quality in use** (CNC-03, CNC-04, CNC-08) and **dependence on a few people** (CNC-06, CNC-07). The indicators are defined in section 12.

These dimensions feed **D3 Data and knowledge** in the maturity model (document 11).

03

Principles

#	Principle	What it implies
1	**No data without an owner**	Every dataset or knowledge source used by an AI system has an identified business owner.
2	**Quality is defined for the use**	Data that are sufficient for a report may be insufficient to train a model. Thresholds are set per use case.
3	**Using data for AI is a purpose**	The legal basis and purpose compatibility are analysed before the data are used, not afterwards.
4	**The minimum necessary**	Only the data and attributes needed are used, for the time needed.
5	**Everything is traceable**	It is possible to reconstruct which data, which model version and which sources produced an output.
6	**AI does not extend permissions**	An AI system does not give access to information that the person using it could not see by other means.
7	**Knowledge expires**	Every knowledge source has a review date and an owner responsible for its validity.
8	**Third-party rights are respected**	Data and content are not used without sufficient title for the specific use.

04

Data governance applied to AI

4.1 Roles

Role	Responsibility in AI
Data owner (business)	Authorises use for a specific use case (at G3), defines acceptable quality and is accountable for the purpose.
Data custodian (technology)	Applies access, quality, retention and backup controls; supports the AI Technical Owner.
AI Technical Owner	Is accountable for the use case's data and models, their preparation and their lineage.

Role	Responsibility in AI
Data protection officer	Advises and monitors in processing of personal data; consulted on the impact assessment (phases 3–4); data protection takes part in the multi-level go-live sign-off at G5.
Content or knowledge owner	Quality, validity and permissions of generative AI sources (document 50, section 4.2).
AI Risk Owner	Assesses RT-DAT risks and issues risk clearance.

4.2 Dataset record for AI

Every dataset used to train, fine-tune, validate, test or feed an AI system in production **must** be catalogued with at least the following fields. If the company already has a data catalogue, the record is implemented as an extension of that catalogue.

Field	Content
Identifier, name, owner and custodian	Principle 1.
Description and origin	What it contains, which systems it comes from, how it was generated; whether it is synthetic (section 8).
Period and volume	Dates covered and number of records.
Personal data	Yes or no; categories, special categories, minors; review by the data protection officer.
Legal basis and authorised purposes	By purpose, with a compatibility analysis if the purpose is new (section 9).
Use and confidentiality restrictions	Contractual, licensing, sectoral, territorial (section 11); classification.
Quality	Dimensions measured, result and date (section 4.3).
Representativeness and known biases	Populations covered and under-represented; mandatory for high risk (section 6).
Retention	Period and rule for deletion or anonymisation (section 4.5).
Systems using it	Use cases and model versions; link to P16 and T02.
Version	Version and freeze date if used for training.

4.3 Quality

Dimension	Question	Example rule
Accuracy	Does it reflect reality?	Sample verified against the primary source.
Completeness	Are values or records missing?	Proportion of nulls per critical attribute.
Consistency	Is it consistent across systems and over time?	Reconciliation of totals between source and target.
Timeliness	Is it up to date for the use?	Maximum age of the latest record.
Validity	Does it comply with format and domain?	Values within permitted lists and ranges.
Uniqueness	Are there duplicates?	Duplicate records by business key.
Representativeness	Does it cover the population and the situations in which the system will be used?	Distribution by segment compared with the population of use.
Label quality	Are the labels used for learning correct?	Inter-annotator agreement on a sample.

Thresholds are set in phase 3 for each critical dataset and become monitoring controls in phase 6 (document 52, section 5).

4.4 Legal basis and minimisation

- The use of personal data to develop or operate an AI system is processing with **its own purpose**. It must have a legal basis (GDPR, Art. 6) and, if the data were collected for another purpose, a **compatibility** analysis (Art. 6(4)) or a new basis.
- **Special categories** of data (Art. 9) require an applicable exception and enhanced controls.
- **Minimisation** (Art. 5(1)(c)) applies to attributes, records and period. In AI, each personal attribute used as a variable is justified in writing.
- **Data protection by design and by default** (Art. 25) is reflected in P15, P16 and P18.
- The processing is included in the **record of processing activities** (Art. 30) and, where applicable, in a **data protection impact assessment** (Art. 35) in P11.

4.5 Retention

Element	Rule
Training, validation and test data	Retained frozen and versioned while the model is in use and for the period needed to reconstruct or audit, within the limits of the legal basis. If they cannot be retained, a description sufficient for traceability is retained.
Input data and outputs in production	According to the log retention period (document 52, section 11) and minimisation.
Prompts and conversations with generative AI	Defined and communicated period; not retained indefinitely by default.
Indexes and vector representations	Follow the period of the source they come from; if the source is deleted, its representation is deleted.
When a system is retired	The treatment of data and models decided at G7 is applied (document 52, section 12).

05

Data preparation by phase

5.1 Requirements by phase

Phase	What is done with the data	Evidence	Stop question
0 · Context	Declare known data constraints: protected data, international transfers, secrets, licences.	P02	Is there a constraint that prevents the use case?
1 · Discovery	Identify what data or knowledge would be needed and whether they exist.	P06	Does the opportunity depend on data that do not exist?
2 · Hypothesis	Confirm that the baseline can be measured with real data.	P09	Can the baseline be measured?
3 · Feasibility	**Availability and quality assessment with real data**; legal basis; third-party rights; representativeness.	P10, P11	Are there sufficient data, of sufficient quality and legally usable?
4 · Design	Lineage, dataset records, quality controls, minimisation, retention, knowledge source permissions.	P15, P16, P18	Does the design treat the data with sufficient controls?

Phase	What is done with the data	Evidence	Stop question
5 · Delivery	Freeze and version datasets; quality, bias and data leakage testing; validate knowledge sources.	P21, P22	Do the results hold with production data?
6 · Operation	Monitor quality, drift and validity; manage data incidents.	P25, P27	Has quality degraded or has the legal basis changed?
7 · Evolution or retirement	Review whether the data allow scaling; treat data and models on retirement.	P30	Do the data support the extended scope?

5.2 Phase 3: the assessment that prevents the most stops

SEVEN-G starts from the premise that **lack of data availability or quality** is the main cause of failure of AI use cases that do not reach production or do not sustain their value. The framework does not base this premise on market figures, but on its logic: a data problem discovered in phase 5 costs the entire build. That is why the assessment is placed in phase 3, before investment in design and build, and indicator DAT-07 makes it possible to test the premise against each company's own data.

Rule. The phase 3 data assessment **must** be carried out **with real data** or with a representative sample extracted from the source systems, not with descriptions of what the data should contain. An assessment based only on interviews or data dictionaries does not pass G3.

Block	Checks
Existence and access	The data exist; authorised access to a sample has been obtained; the cost and timeframe of full access are known.
Volume and history	There are sufficient records and period for the type of system; there are sufficient cases of the relevant situations, including minority ones.
Measured quality	Results by dimension (section 4.3) on the sample, against proposed thresholds.
Labels or ground truth	There is a way of knowing whether the system is right: historical labels, expert review or observable outcome, and with what delay.
Representativeness	The sample covers the population and the contexts of use; known biases documented.
Legitimacy	Legal basis and purpose; compatibility analysis; special categories; transfers; licences and third-party rights.
Availability in production	The data used for building will be available in production with the same definition, frequency and latency.
Knowledge (generative AI)	Sources identified, with known owner, validity and permissions; proportion of outdated content.

Block	Checks
Preparation cost	Estimated effort for cleansing, integration, labelling and curation, included in the full cost (document 42).

Result. Each block is rated as *sufficient*, *sufficient with actions* (action, cost, timeframe and owner) or *insufficient*. An *insufficient* legitimacy block prevents proceeding. If other blocks are insufficient, the expected outcome of G3 is **Iterate**, **Pivot** or **Stop** with the coded reason *Insufficient data* (document 03, section 3.3).

06

High-risk systems: Article 10 of the EU AI Act

This section summarises, without legal detail, what Article 10 of the EU AI Act requires of the training, validation and test datasets of **high-risk systems**, and where it is reflected in SEVEN-G. The obligation-by-obligation analysis, the application dates (Annex III from 2 December 2027 and Annex I from 2 August 2028, following Regulation (EU) 2026/1744) and the other amendments are dealt with in document 34 (§3.1 and §3.2).

Requirement in summary	Where it is reflected in SEVEN-G
Data governance and management practices appropriate to the intended purpose: design choices, origin and original purpose of the data, preparation (annotation, labelling, cleaning, updating, enrichment, aggregation), assumptions about what the data measure and represent.	Dataset record (4.2); P16.
Prior assessment of the availability, quantity and suitability of the datasets needed.	Phase 3 assessment (5.2); P10.
Examination of possible biases likely to affect health, safety or fundamental rights or lead to discrimination, and measures to detect, prevent and mitigate them.	P22 (bias testing); document 50, section 8.4; P12.
Identification of data gaps or shortcomings and how they can be addressed.	Phase 3 assessment; recorded actions.
Datasets that are relevant, sufficiently representative and, to the best extent possible, free of errors and complete in view of the intended purpose.	Quality (4.3) with thresholds per use case.
Consideration of the characteristics of the geographical, behavioural, contextual or functional setting of use.	Representativeness and availability in production blocks (5.2).

Requirement in summary	Where it is reflected in SEVEN-G
Exceptional processing of special categories of personal data to detect and correct biases, subject to strict conditions and safeguards.	Section 9.4; decision documented with the data protection officer.

When the company acts as the **deployer** of a supplier's high-risk system, it must ensure that the **input data** under its control are relevant and sufficiently representative in view of the intended purpose (Art. 26(4)), and require the necessary information from the supplier (document 36).

This document does not constitute legal advice.

07

Data and model lineage (P16)

7.1 What it is and what it is for

Lineage makes it possible to answer, for any output of an AI system: **which data, which transformations, which model version, which configuration and which knowledge sources produced it**. It is the basis for investigating incidents, handling individuals' rights, auditing, reproducing results, managing changes and retiring systems.

7.2 Minimum content

Element	Content	Lite	Enterprise
Sources	Source systems, datasets (with record) and knowledge sources.	Yes	Yes
Transformations	Preparation steps, rules, versioned code.	Summary	Detail with version
Training, validation and test sets	Frozen version, date, split criterion.	If there is in-house training	Yes
Model	Type, version, configuration parameters, training date, validation results.	Yes	Yes
Third-party model	Supplier, exact version identifier, adoption date, data use terms.	Yes	Yes

Element	Content	Lite	Enterprise
System prompts and templates (generative AI and agents)	Prompt version, available tools, filters.	Yes	Yes
Retrieval indexes	Indexed sources, indexing date, chunking and retrieval parameters.	If there is retrieval	Yes
Consumers	Processes, systems and decisions that use the outputs.	Yes	Yes
Link to logs	Which identifiers make it possible to link a specific output to the above versions.	Recommended	Yes

7.3 Rules

1. P16 is prepared in phase 4 and **must** be current for G5.
2. Any change in sources, transformations, model, prompts or indexes **updates** P16 and is recorded in P27 (document 52, section 6).
3. Lineage is generated **automatically** wherever the platform allows; manual documentation is limited to what cannot be captured.
4. A system in production without current lineage is a **nonconformity** (minor if it does not affect decisions; major if the system is Enterprise).

08

Synthetic data

8.1 When they make sense

Use	Illustrative example	Value
Supplementing scarce cases	Infrequent situations that the model must recognise.	Improved robustness.
Testing and development	Test environments without real personal data.	Lower exposure.
Stress and security testing	Adversarial cases to evaluate an assistant or an agent.	Better evaluation.
Sharing with third parties	Suppliers that develop without access to real data.	Lower contractual and privacy risk.

8.2 Risks and controls

Risk	Control
False sense of anonymity: synthetic data generated from personal data may allow re-identification of, or inference of information about, real people.	Re-identification risk assessment before treating them as non-personal; until this is demonstrated, they are treated as personal data.
Loss of fidelity: they do not reproduce relevant real relationships.	Utility validation: comparison of distributions and of the performance of models trained with real versus synthetic data.
Amplification of biases in the source dataset.	Bias testing on the synthetic dataset and on the resulting model.
Circular validation: validating a model with data generated by the same process that trained it.	Final validation **must** be carried out with real data or with an independent real sample.
Contamination: unidentified mixing of real and synthetic data.	Mandatory marking of synthetic origin in the record and in the lineage.

8.3 Rules

1. Every synthetic dataset has **its own record** (4.2) indicating the source dataset, the generation method and its version.
2. No G5 validation result is based **solely** on synthetic data.
3. Synthetic data in use have a recorded utility and privacy validation (DAT-11).

09

Personal data in training, prompts and outputs

9.1 In training and fine-tuning

Issue	What SEVEN-G requires
Purpose and legal basis	Documented in the record and verified by the data protection officer before G3.
Anonymisation versus pseudonymisation	Only effectively anonymised data fall outside the GDPR. Pseudonymisation reduces risk, but the data remain personal. Anonymisation is justified with a re-identification risk analysis.

Issue	What SEVEN-G requires
Memorisation	Models trained with personal data may reproduce them in their outputs. Data extraction is tested in phase 5 when the model is trained or fine-tuned with personal data. The European Data Protection Board has analysed in its Opinion 28/2024 when an AI model can be considered anonymous; consulting it is recommended.
Data subjects' rights	It is documented how access, rectification, erasure and objection are handled when the data have been used for training, and what this implies for the model (retraining, output filters or other measures).
Transparency	Information to data subjects (GDPR, Arts. 13 and 14) covers use for AI when it is a new purpose.

9.2 In prompts and context

The prompts that employees write in generative AI tools, the documents they attach and the context that is retrieved automatically are **data that leave the direct control** of the person and, often, of the company.

Risk	Control
Employees entering personal or confidential data into unauthorised tools.	Acceptable use policy with a classification of the information permitted per tool (document 31); data leakage prevention controls; usage monitoring (T21).
Supplier retaining prompts or using them to train its models.	Contractual terms verified in P14: no use for training, retention period, location, sub-processors (document 36).
Automatic retrieval of personal data not needed for the response.	Minimisation in the retrieval design; personal data filters on input and output.
System prompts containing sensitive information.	System prompts do not contain secrets or personal data; they are reviewed in P18.
Conversation logs retained without a time limit.	Defined period (4.5) and restricted access.

9.3 In outputs

AI systems may **infer** personal data that nobody has provided (for example, traits, preferences or sensitive situations). An inference about an identified person is personal data. If the inference relates to a special category, its rules apply. The design (P17) **must** identify which inferences the system produces and whether they are necessary and lawful.

9.4 Data for detecting bias

Detecting bias may require data on protected characteristics that the company does not normally process. For high-risk systems, the EU AI Act exceptionally allows the processing of special categories for that purpose and with strict safeguards (Art. 10(5)), and Regulation (EU) 2026/1744 extends that possibility beyond high risk subject to a strict necessity criterion (document 34 §3.2). In all cases, the decision is taken with the data protection officer and the legal analysis, documented in P11, and measures such as separation, pseudonymisation, restricted access and deletion at the end of the analysis are applied.

10

Knowledge

10.1 Knowledge sources for generative AI

When a generative AI system responds on the basis of the company's documents and content through retrieval (*RAG*, retrieval-augmented generation), **the quality of the responses depends more on the sources than on the model**. SEVEN-G treats each source as a governed asset.

Source inventory. Each system maintains the list of sources it uses, with:

Field	Content
Source	Repository, space, knowledge base or system.
Content owner	Person or function accountable for quality and validity.
Scope	Which parts are indexed and which are expressly excluded.
Content type	Internal rules, procedure, product, technical, legal, commercial, historical.
Authority	Official, supporting or informative source; what prevails in the event of contradiction.
Classification and permissions	Confidentiality and source permissions model.
Validity	Mandatory review date by content type.
Index update frequency	Continuous, daily, weekly.
Rights	Whether it includes licensed third-party content (section 11).

10.2 Admission criteria

A source **should not** be connected to a generative AI system if:

- It has no content owner.
- It cannot be determined which documents are current and which are not.
- It mixes drafts, old versions and official documents without distinguishing them.
- Its permissions model cannot be carried over to retrieval (section 10.4).
- It includes personal data or third-party content whose use has not been analysed.

10.3 Curation and validity

Activity	What is done	Frequency
Initial clean-up	Remove duplicates, superseded versions and drafts; mark the current document.	Before G5
Minimum metadata	Owner, publication date, review date, status (current, superseded, historical), authority.	When each document is added
Validity review	The content owner confirms, updates or withdraws documents that reach their review date.	By type; indicatively, 6 months for procedures and 12 for stable content
Removal from the index	Superseded or expired documents are removed from the index or marked as historical and excluded from retrieval by default.	At each update
Gap detection	Analysis of queries with no relevant document retrieved (CNC-08) to create or improve content.	Monthly
Contradiction detection	Review of documents that give incompatible answers on the same topic.	Quarterly

Rule. A system that cites an expired document in a response on terms, prices, rules or procedures affecting customers or employees generates an **incident** that is classified using the S1–S4 scale (document 52, section 7).

10.4 Access permissions

Content retrieval is a **route of access to information** and must respect the same permissions as the source.

Risk	Control
Overexposure: the assistant shows a person content that they could not open in the source.	Filtering by user permissions **at retrieval time**, not only at indexing; specific testing in phase 5.
Excessive inherited permissions in the source, which AI makes visible.	Review of source permissions before connecting it; AI makes it easy to find what was previously "hidden by obscurity".
Common index for groups with different permissions.	Separate indexes or verified filtering by group.
Agents with their own permissions that retrieve information for users with fewer permissions.	Agents act with the permissions of the user they serve, unless there is a justified design in P18 (document 35).
Permission changes in the source not carried over to the index.	Permission synchronisation with a defined and monitored maximum time limit.

10.5 Assessment of the quality of knowledge in use

Aspect	What is measured	Indicator
Retrieval relevance	Whether the documents retrieved are appropriate for the query.	Evaluated on the evaluation set (document 52, section 4.2).
Faithfulness to sources	Whether the statements in the response are supported by the retrieved documents.	CNC-04
Use of expired content	Whether documents that are no longer valid are cited.	CNC-03
Gaps	Queries without a relevant document.	CNC-08
Source citation	Whether the response allows the user to check the source.	Design requirement in P17.

10.6 Critical knowledge dependent on a few people

Step	Activity	Result
1 · Identify critical domains	Processes, products, systems or relationships whose knowledge is essential to operate or decide.	List of critical domains (C1).
2 · Measure dependence	For each domain: how many people master it, whether it is documented and usable, and departure risk (retirement, turnover, mobility).	Dependence map (CNC-06).
3 · Prioritise	Domains with two people or fewer, without usable documentation and with departure risk within the thesis horizon.	Capture plan.
4 · Capture	Structured interviews (which may be AI-assisted), recording of explanations, case reviews, documentation of decisions and exceptions.	Content with owner and validity.
5 · Validate	The expert and a second reviewer confirm that the content is correct and complete for its use.	Validated content.
6 · Make available	Incorporation as a governed source (section 10.1).	CNC-07

Considerations regarding people. The capture of tacit knowledge is carried out with **informed voluntary participation**, recognition of the expert and clarity about the subsequent use of the content. It **must not** be presented or used as a step prior to dispensing with the person without the effect having been assessed in accordance with document 50. If recordings or transcripts contain personal data, the corresponding legal basis, information and retention apply.

11

Intellectual property and third-party rights

Consultation carried out in September 2026. This is an area of evolving legislation and case law; it must be verified with specialist advice.

11.1 Input data and content

Situation	Issue	Control
Acquired or licensed data	The licence may not permit use for training or feeding AI systems, or use by suppliers.	Licence review before G3; restrictions in the record (4.2).

Situation	Issue	Control
Content obtained from public sources on the internet	The fact that content is accessible does not mean its use is free. In the European Union, text and data mining is subject to exceptions with conditions, and rightholders may reserve that use for purposes other than scientific research (Directive (EU) 2019/790 , Arts. 3 and 4).	Prior legal analysis; respect for rights reservations expressed in an appropriate manner; source record.
Customer or partner content	Its use may be limited by contract and by confidentiality.	Contractual verification; excluded by default.
The company's own trade secrets	Entering them into third-party services may compromise the reasonable protection measures required for them to remain secret (Directive (EU) 2016/943 ; in Spain, Ley 1/2019 de Secretos Empresariales (Law 1/2019 on Trade Secrets)).	Classification, authorised tools and contractual terms.
Open-source software and models	Licences may impose conditions of use, attribution or purpose restrictions.	Licence record in P15; review before G4.

11.2 Third-party general-purpose models

The EU AI Act requires **providers of general-purpose AI models** to put in place a policy to comply with copyright law, including rights reservations, and to publish a sufficiently detailed summary of the content used for training (Art. 53(1)(c) and (d)). A company that uses those models **should** request and retain that information in the supplier assessment (P14) and assess the contractual safeguards against third-party claims (document 36).

11.3 Generated outputs

Issue	What SEVEN-G requires
Ownership	Copyright protection of content generated with AI is uncertain and depends on the jurisdiction and the degree of human contribution. Content that the company needs to protect is generated with documented human contribution or protected by other means (contract, secrecy).
Reproduction of third-party works	Outputs may reproduce protected content. In external uses (marketing, product, publications), human review and, where appropriate, detection tools are applied.
Supplier terms	It is verified what rights the supplier grants over the outputs and what use the supplier may make of them.
Transparency	Any applicable obligations to mark or disclose generated or manipulated content are met (document 34).

This document does not constitute legal advice.

12

Indicators

The codes are provisional; their definitive `IND-` code is shown in the mapping table in document 41 (§21). Knowledge indicators use the **CNC-** prefix so that they are not confused with the `IND-CON` family (board) in document 41. Thresholds are **indicative and to be calibrated**. "No data" is not zero.

12.1 Sphere 05 · Data

Code	Indicator	Formula	Frequency	Source	Indicative reference
DAT-01	Ownership coverage	Datasets used by systems in production with an assigned owner ÷ datasets used by systems in production	Quarterly	Catalogue, T02	100%
DAT-02	Catalogue coverage	Datasets used by AI with a complete record ÷ datasets used by AI	Quarterly	Catalogue	100% in Enterprise
DAT-03	Legal basis coverage	Datasets with personal data used by AI with verified legal basis and purpose ÷ datasets with personal data used by AI	Quarterly	Catalogue, record of processing activities	100%
DAT-04	Quality by dimension	Records that pass the dimension's rules ÷ records assessed (by dataset and dimension)	According to load frequency	Quality monitoring	Threshold set per use case in phase 3
DAT-05	Critical datasets within threshold	Critical datasets that meet all their thresholds ÷ critical datasets	Monthly	Quality monitoring	≥ 95%
DAT-06	Lineage coverage	Systems in production with current P16 ÷ systems in production	Quarterly	T02, P16	100%
DAT-07	G3 decisions due to data	G3 decisions with an Iterate, Pivot or Stop outcome and the reason *Insufficient data* ÷ G3 decisions	Half-yearly	T01, T03	For interpretation, not a target: very low may indicate a lax assessment; very high, poor opportunity selection

Code	Indicator	Formula	Frequency	Source	Indicative reference
DAT-08	Retention compliance	Datasets with a defined and applied retention period ÷ datasets used by AI	Half-yearly	Catalogue	100%
DAT-09	Incidents caused by data	AI incidents whose root cause lies in data ÷ closed AI incidents	Quarterly	T08	Decreasing trend
DAT-10	Data access time	Median days between a data access request for a use case and its authorised availability	Quarterly	Access requests	Decreasing trend
DAT-11	Validated synthetic data	Synthetic datasets in use with recorded utility and privacy validation ÷ synthetic datasets in use	Half-yearly	Catalogue	100%

12.2 Sphere 06 · Knowledge

Code	Indicator	Formula	Frequency	Source	Indicative reference
CNC-01	Sources with an owner	Knowledge sources connected to AI with a content owner ÷ connected sources	Quarterly	Source inventory	100%
CNC-02	Content validity	Indexed documents within their review date ÷ indexed documents	Monthly	Source metadata	≥ 95%
CNC-03	Use of expired content	Sampled responses citing documents that are no longer valid ÷ sampled responses with a citation	Monthly	Continuous evaluation	Close to 0%
CNC-04	Faithfulness to sources	Evaluated responses whose statements are supported by the retrieved sources ÷ evaluated responses	Continuous or monthly	Continuous evaluation	Threshold per use case; shared with OPE-09
CNC-05	Improper exposure	Findings of access to content without permission detected in testing or in production, per period	Quarterly	Permission testing, T08	0

Code	Indicator	Formula	Frequency	Source	Indicative reference
CNC-06	Critical knowledge dependence	Critical domains mastered by two people or fewer and without usable documentation ÷ critical domains	Annual (C1, C5)	Dependence map	Decreasing trend
CNC-07	Capture coverage	Dependent critical domains with captured and validated content ÷ dependent critical domains	Half-yearly	Capture plan	As per plan
CNC-08	Queries without a relevant source	Queries with no relevant document retrieved ÷ evaluated queries	Monthly	Continuous evaluation	Decreasing trend

13

Common mistakes

#	Mistake	Consequence	How it is avoided
1	Assessing data in phase 3 using descriptions rather than real data.	It is discovered in phase 5 that the data are not fit for purpose.	Section 5.2.
2	Assuming that a system's data can be used for any purpose.	Processing without a legal basis.	Legitimacy as a blocking block at G3.
3	Building with data that will not be the same in production.	Actual performance far below the validated performance.	Availability in production block.
4	Having no way of knowing whether the system is right.	Impossible to validate and monitor.	Labels block.
5	Connecting an entire repository to an assistant.	Drafts, old versions, overexposure.	Sections 10.2 and 10.3.
6	Filtering permissions at indexing rather than at retrieval.	Employees see what they should not.	Section 10.4.
7	Treating pseudonymised or synthetic data as anonymous.	Breach of the GDPR.	Sections 8.2 and 9.1.
8	Not versioning training datasets.	Impossible to reproduce or investigate.	Sections 4.2 and 7.

#	Mistake	Consequence	How it is avoided
9	Not updating the lineage.	False lineage when an incident occurs.	Section 7.3.
Sphere map and ambition levels (10)	Using third-party content because it is on the internet.	Intellectual property claims.	Section 11.1.
Maturity model (11)	Omitting the cost of preparation, labelling and curation.	Inflated expected net value at G3.	Section 5.2; document 42.
Transformation index (12)	Capturing experts' knowledge without their informed participation.	Poor content and loss of trust.	Section 10.6; document 50.

14

Typical data and knowledge risks

Codes and names from the typical risk catalogue in document 33; two of these risks appear there in other categories (LEG, ORG).

Code (document 33)	Risk	Main controls
RT-DAT-01	Insufficient quality: insufficient data or data of inadequate quality for the purpose.	Phase 3 assessment with real data; thresholds and monitoring.
RT-DAT-03	Lack of legal basis or incompatible purpose in the use of personal data.	Record, verification by the data protection officer, impact assessment.
RT-DAT-02	Non-representative data and bias that produce discriminatory outcomes.	Representativeness, bias testing, section 9.4.
RT-DAT-04	Unknown lineage: loss of traceability of data and models.	P16 current and updated with every change.
RT-DAT-05	Data exposure in the AI lifecycle: leakage of personal or confidential data through prompts, retrieval or memorisation.	Sections 9.2 and 10.4; extraction testing.
RT-DAT-06	Obsolete or contradictory knowledge in responses.	Curation and validity; CNC-02 and CNC-03.
RT-LEG-06	Infringement of intellectual property or other third-party rights through the data or content used.	Section 11; supplier assessment.

Code (document 33)	Risk	Main controls
RT-ORG-02	Dependence on key people: loss of critical knowledge when people leave.	Section 10.6; CNC-06 and CNC-07.

15

Associated tools and templates

This document does not create a new tool. The dataset record and the knowledge source inventory are implemented in the **data catalogue that the company already uses** or, if none exists, in template P64 and as system attributes in T02.

Name	Use in this document
Inventory, *gates*, risks, regulatory classifier (T02 · T03 · T06 · T07)	Link between systems, data and lineage; G3 assessment; RT-DAT risks; high-risk obligations.
Incidents, suppliers, corporate use (T08 · T09 · T21)	Incidents caused by data; supplier terms; data in unauthorised tools.
Context and constraints statement (P02)	Phase 0 data constraints.
Feasibility assessment (P10)	Data availability and quality assessment (section 5.2).
Regulatory classification and impact assessments (P11)	Legal basis, data protection impact assessment, data for detecting bias.
AI supplier assessment (P14)	Use of prompts and data by the supplier; copyright policy of general-purpose models.
Architecture record (P15)	Licences for components and models.
Data and model lineage (P16)	Main template of this document; uses the records and the source inventory in P64.
Security design (P18)	Retrieval and agent permissions.
Validation and pilot results (P22)	Quality, bias, data leakage and permission testing.
Monitoring and alert configuration (P25)	Data quality and validity thresholds.

Name	Use in this document
Dataset record and knowledge source inventory (P64)	Record of each dataset and inventory of knowledge sources, if the company has no data catalogue.

16

Related documents

Document	Relationship
00, 01 and 03 · Foundations and register	Spheres 05 and 06; data activities by phase; *Insufficient data* reason.
10, 11 and 20 · Spheres, maturity and phase manuals	Spheres 05 and 06; dimension D3; requirements of section 5 in each phase.
31 · Corporate policy and acceptable use	Information that may be entered into each tool.
33 · AI risk methodology	Consolidation of RT-DAT risks.
34 · Regulatory mapping	Article 10 of the EU AI Act , GDPR and intellectual property in detail.
35 · AI and agent security	Protection of data against attacks; agent permissions.
36 · AI third parties and suppliers	Clauses on data use, confidentiality and intellectual property.
41 · Indicator catalogue	Consolidation of the DAT and CNC indicators (mapping table in 41 §21).
42 · AI costs	Cost of preparation, labelling and curation (*data* category).
50 · People and organisation	Capture of expert knowledge and employee data.
52 · AI operations manual	Quality and drift monitoring, logs, changes and retirement.

17

Version control

Version	Date	Changes
0.1	16-09-2026	First version. Develops spheres 05 and 06 with indicators; data governance applied to AI (roles, dataset record, quality, legal basis, minimisation and retention); data requirements by phase with the phase 3 assessment using real data; the summary of Article 10 of the EU AI Act ; the content of P16; synthetic data; personal data in training, prompts and outputs; knowledge sources for generative AI, curation, validity, permissions and critical knowledge dependent on a few people; intellectual property; nineteen indicators with a formula, twelve common mistakes and eight typical risks. Consistency adjustments with 01 (segregation of duties in Lite, R6 outcomes, agents criterion) and with 34 and 37; risks aligned with the codes in document 33. Knowledge indicators renamed from CON- to CNC- to avoid the clash with the IND-CON family (board) in document 41.

External sources cited

Links to the official or primary source, checked before publication. Always verify the version in force on the date of use.

Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure · https://eur-lex.europa.eu/eli/dir/2016/943/oj/eng · Parlamento Europeo y Consejo de la UE · 08-06-2016

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market · https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng · Parlamento Europeo y Consejo de la UE · 17-04-2019

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) · https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng · Parlamento Europeo y Consejo de la UE · 27-04-2016

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) · https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng · Parlamento Europeo y Consejo de la UE · 13-06-2024

Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital

Omnibus on AI) · https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng · Parlamento Europeo y Consejo de la UE · 08-07-2026

European Data Protection Board (EDPB) · https://www.edpb.europa.eu/edpb_en · Comité Europeo de Protección de Datos (organismo de la UE creado por el art. 68 del RGPD)

Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models · https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en · Comité Europeo de Protección de Datos · 17-12-2024