Sonae

30th October 2026

EDPB Guidelines on Web Scraping

Sonae welcomes the opportunity to contribute to the public consultation launched by the European Data Protection Board (EDPB) on its draft Guidelines 03/2026 on web scraping in the context of generative AI. Given the growing use of these techniques in the development of generative AI solutions, the EDPB's effort to clarify their legal and technical implications under the GDPR is a timely initiative.

From Sonae's perspective, the relevance of these Guidelines goes beyond web scraping itself. They address concepts of broader relevance, such as the allocation of responsibilities, legal bases and the application of data protection principles, on which much of the existing guidance dates back several years.

Sonae therefore welcomes the clarification that the public accessibility of personal data on the internet does not mean that it ceases to be personal data, nor that data subjects have consented to its use for training AI models.

However, several aspects of the Guidelines would benefit from further clarification to ensure legal certainty, notably regarding the incidental collection of special categories of data, the exercise of data subjects' rights after training, the due diligence required when acquiring datasets, transparency obligations and the allocation of responsibilities between entities.

Against this background, Sonae's position is structured around the following key messages:

1. Clearer criteria are needed for the incidental collection of special categories of data

The Guidelines, in paragraph 68, recognise that even where measures are in place to prevent it, it is not always possible to assess with reasonable certainty whether scraping leads to the collection of special categories of personal data until after the data has been scraped.

In this context, the EDPB considers that the reasoning of the Court of Justice in GC and Others (C-136/17), concerning search engines, can be relevant for the incidental and residual collection of such data, so that the prohibition in Article 9(1) GDPR applies within the framework of the controller's responsibilities, powers and capabilities, as disposed in paragraphs 72 to 74. This is not a general exemption. It depends on the activity having relevant similarities with that of a search engine, on the collection being only incidental and residual, on it being difficult or impossible to detect such data in advance, and on the implementation of measures to prevent their collection, retention and dissemination.

However, beyond the general definition in footnote 44, the Guidelines do not specify when collection can be considered merely incidental and residual, notably in terms of the volume or proportion of the data, the foreseeability of the collection and the minimum effectiveness of the preventive measures.

In addition, it is unclear what the criterion of "responsibilities, powers and capabilities" means in practice when scraping is carried out by a processor or when the dataset is acquired from a third party.

Therefore, the Guidelines should:

- **Specify the criteria to determine when the collection of special categories of personal data can be considered incidental and residual, notably regarding the volume or proportion of the data, the foreseeability of the collection and the minimum effectiveness of preventive measures;**
- **Clarify how the criterion of "responsibilities, powers and capabilities" applies where scraping is carried out by a processor or where the dataset is acquired from a third party.**

2. Practical guidance is needed on the exercise of data subjects' rights after training

The Guidelines acknowledge that data subjects may find it difficult to exercise their rights and to oppose the scraping of their data, notably when it is carried out by third parties. They also recognise, in paragraph 60, that given the current state of the art, personal data cannot be easily deleted from a model once it has been trained.

As mitigating measures, the Guidelines recommend facilitating the exercise of rights, providing opt-out mechanisms, including before the collection takes place, and deleting or anonymising unnecessary personal data as soon as possible, as in paragraph 66. They also refer to output filters and, in the longer term, to model unlearning, as described in paragraph 74.

However, it remains unclear how rights such as erasure and objection should be ensured where the data has already been used for training and cannot be isolated or removed from the model, and what role alternative measures, such as output filters and prior opt-out, play in that context. Clarification on this point would be relevant well beyond the scope of these Guidelines.

Therefore, the Guidelines should:

- **Clarify how data subjects' rights, such as erasure and objection, should be ensured where the data has already been used for training and cannot be isolated or removed from the model;**
- **Clarify the role of alternative measures, such as output filters and prior opt-out mechanisms, in ensuring the exercise of those rights.**

3. The level of due diligence required when acquiring third-party datasets should be defined

The Guidelines distinguish between situations in which an organisation contracts a provider to carry out the scraping on its behalf and situations in which it acquires a dataset previously collected by another entity for its own purposes.

In the first case, where the provider collects the data in accordance with the documented instructions of the AI developer, it may act as a processor, with the developer determining the purposes and essential means of the collection (paragraph 18). In the second case, the provider and the organisation reusing the dataset are, in principle, responsible for their own, separate processing activities (paragraph 19). The reusing organisation must therefore identify a legal basis for its own processing and assess the compliance of the datasets and sources it intends to use.

However, the Guidelines do not set out a minimum level of diligence or evidence that the acquirer should obtain to assess the provenance and lawfulness of the data. This is particularly relevant when groups of companies contract AI providers.

Therefore, the Guidelines should:

- **Define the level of due diligence required from organisations acquiring or reusing datasets previously collected by third parties, regarding the provenance and lawfulness of the collection;**
- **Specify the minimum evidence and contractual guarantees that acquirers should obtain.**

4. Transparency obligations under Article 14 GDPR should be workable in practice

Where informing data subjects individually proves impossible or would require disproportionate effort, the controller must make the information publicly available, including a precise indication of the sources of the data. Where a complete list of sources cannot be provided, sources should be listed to the greatest extent possible, together with an indication of the types of sources not included and the reasons for their exclusion, as in paragraph 30. It is also considered good practice to include domain names and URLs in a searchable format, as well as the date or period of collection, as proposed by paragraph 31.

However, it is unclear what level of detail is acceptable in this public information, in particular regarding the list of sources where enumerating all of them would be disproportionate. It is also unclear whether a single publication may serve several projects of the same controller.

Therefore, the Guidelines should:

- **Clarify the acceptable level of detail of the publicly available information, regarding the list of sources where enumerating all of them would be disproportionate;**
- **Confirm that a single publication may serve several projects of the same controller.**

5. The allocation of responsibilities in multi-party and group structures should be clarified

The Guidelines state that the roles of the organisations involved should be assessed on a case-by-case basis, as mentioned in paragraph 16. They provide general considerations on when the scraping provider acts as a processor, when the parties are separate controllers and when they may be joint controllers, because they determine the purposes and means of the processing together (paragraphs 18, 19, 20 and Example 1).

However, in projects involving several parties, and in particular in corporate groups where several entities cooperate in the development or use of AI models, it is unclear how to distinguish joint controllership from separate controllership where one entity carries out the scraping at the request of another, and other entities develop and use the model.

Therefore, the Guidelines should:

- **Clarify the criteria for distinguishing joint controllership from separate controllership where one entity carries out the scraping at the request of another and other entities develop and use the model, in particular within corporate groups.**

Conclusion

The draft Guidelines provide an important clarification of how the GDPR applies to web scraping in the context of generative AI, at a time when these techniques are increasingly used.

From Sonae's perspective, their relevance goes beyond web scraping itself. The concepts they address, namely the allocation of responsibilities, the treatment of special categories of data, the exercise of data subjects' rights and transparency, are relevant to any organisation that develops, acquires or uses AI solutions.

These questions are particularly relevant for groups of companies, in which several entities may cooperate in the development or use of AI solutions and rely on third-party providers. Clear guidance on the due diligence required when acquiring datasets and on the allocation of responsibilities between entities would therefore be valuable, both for Sonae and for many other European groups.

In this sense, clear and workable guidance on these matters would strengthen legal certainty, protect data subjects and support the responsible use of AI across the European economy.

Contacts:

Frederico Mota
Head of Public Affairs | Sonae
fpmota@sonae.pt | publicaffairs@sonae.pt

About Sonae:

Sonae is a Portuguese multinational and an active investment holding company, with a diversified portfolio of leading businesses in sectors such as retail, telecommunications, financial services, technology, property and new ingredients. Operating in around 60 countries, it is the largest private employer in Portugal and achieved a turnover of 11.4 billion euros in 2025. Sonae's commitment to sustainability has been recognised internationally, notably through its inclusion in the S&P Sustainability Yearbook 2026 and on the Carbon Disclosure Project (CDP), which distinguishes around 1% of the more than 23,000 companies assessed globally.

Sonae is registered in the European Union Transparency Register under number 687748032165-82.

Find out more at **www.sonae.pt**