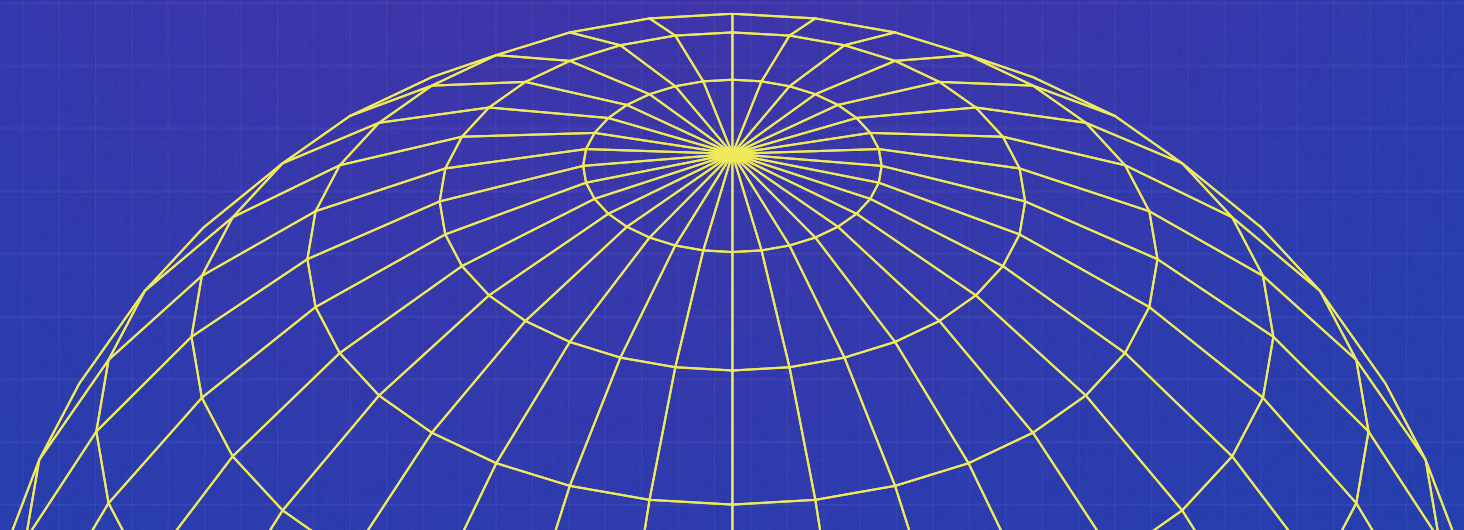
DOT europe

October 2026

DOT EUROPE RESPONSE TO CONSULTATION

DOT Europe contribution to the EDPB consultation on Guidelines 03/2026 on web scraping in the context of generative AI



DOT Europe contribution to the EDPB consultation on Guidelines 03/2026 on web scraping in the context of generative AI

1. Introduction

DOT Europe welcomes the opportunity to comment on the EDPB draft Guidelines 03/2026 on web scraping in the context of generative AI (the "draft Guidelines").[1] DOT Europe represents a broad range of companies operating in Europe and worldwide including the largest companies developing and deploying generative AI in Europe. We, therefore, share the EDPB's objective of clear, practical guidance that supports both innovation and a high level of data protection.

We strongly welcome that the EDPB issues guidance on the topic, and we particularly welcome that the guidelines usefully confirm that web scraping for AI training can rely on legitimate interest, and open a pragmatic route for the incidental processing of special categories of data. In several places, however, we note that recommendations are framed as fixed, cumulative requirements that do not reflect how general-purpose models are built. Our comments aim to make the guidance workable in practice without lowering the protection it offers individuals.

The timing is also key. The co-legislators are currently negotiating internally to find their respective positions on the Digital Omnibus, which would confirm that the development and operation of AI may rely on legitimate interest (proposed Article 88c GDPR) and introduce a derogation for residual special categories of data (proposed Articles 9(2)(k) and 9(5) GDPR).[2] We encourage the EDPB to keep the final Guidelines flexible and risk-based, so that they remain consistent with the outcome of those negotiations and with the AI Act. We would also recommend giving stakeholders the opportunity to comment on any revised text once the Digital Omnibus legislative process concludes.

2. What we welcome

DOT Europe particularly welcomes the following elements:

[1] EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted on 7 July 2026 for public consultation.

[2] European Commission, Proposal for a Digital Omnibus Regulation, COM(2025) 837 final, 19 November 2025; EDPB-EDPS Joint Opinion 2/2026 on the Digital Omnibus, 10 February 2026. At the time of writing, the GDPR-related provisions remain under negotiation in the Council and the European Parliament.

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

- **Legitimate interest (paras 43, 46-50).** The confirmation that Article 6(1)(f) GDPR is an appropriate legal basis for scraping for generative AI training. We also welcome para 50, which allows the development of a general-purpose model to be assessed by reference to the objective pursued, even before its precise uses are decided.
- **Data minimisation and large datasets (para 36).** The clarification that data minimisation does not prohibit training on large volumes of data.
- **Transparency (paras 25, 28 and 30).** The recognition that the disproportionate effort exception in Article 14(5)(b) GDPR can apply to large-scale scraping, that the assessment is made on the dataset as a whole, and that public information is the appropriate way to inform individuals.
- **Article 11 GDPR (para 29).** The confirmation that controllers need not acquire additional information to identify individuals in unstructured scraped data solely to handle data subject requests.
- **Reasonable expectations (para 64 and Example 4).** The acknowledgement that people may be aware that data they publish online can be accessed and reused by third parties. Web crawling has long underpinned the open internet, including search engine indexing, and public awareness of generative AI is high and growing.
- **Lifecycle safeguards (paras 66(i), 76 and 77).** The recognition that measures against memorisation and regurgitation count as mitigation, that measures are non-exhaustive and case-by-case, and that they should evolve with the state of the art.
- **Special categories of data (paras 72–74).** The application of the CJEU's *GC and Others* (C-136/17) ruling to the incidental and residual collection of special categories of data. Assessing the controller's obligations "within the framework of its responsibilities, powers and capabilities" is the right framework.[3]

3. Overarching principle: proportionate safeguards across the AI lifecycle

The GDPR is technology neutral (Recital 15). It requires controllers to achieve outcomes, such as lawfulness, minimisation and effective protection of rights, rather than to use specific tools. The draft Guidelines recognise this where they state that measures are non-exhaustive and must be assessed case by case (para 66), and that they should follow the state of the art (para 77).

Elsewhere, however, and in particular in paras 37, 38, 42 and 74, measures are listed as steps the controller "should" take, often in a fixed sequence. Combined with the accountability requirement in para 75, these lists risk being read as cumulative checklists. The effective safeguards depend on the purpose of the crawl, the type of model and the risks involved.

We therefore invite the EDPB to apply three principles throughout the final Guidelines:

[3]CJEU, C-136/17, *GC and Others v CNIL*, 24 September 2019, ECLI:EU:C:2019:773. See also Morrison Foerster, *Legal analysis of the Digital Omnibus proposal amending Article 9 GDPR*, memorandum to DOT Europe, 10 March 2026 ("Legal Assessment"), paras 17 and 33.

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

- **A toolbox, not a checklist.** Present the measures in paras 37, 38, 42 and 74 as options from which controllers select an appropriate combination for their processing.
- **A holistic, lifecycle assessment.** Assess safeguards across source selection, pre-processing, training, output controls and rights handling together, so that strong measures at one stage can compensate for technical limits at another.
- **Reasonable steps, not guarantees.** The standard should be genuine, documented and regularly reviewed effort proportionate to the state of the art, not a guarantee that no personal or sensitive data is ever processed.

This approach mirrors the continuous, lifecycle-based risk management of the AI Act. It is also the approach the EDPB and EDPS recommended for the Digital Omnibus in their Joint Opinion 2/2026.[4] Finally, it keeps the Guidelines relevant as technology develops.

4. Specific concerns and recommendations

4.1 Broad and diverse data is a condition for safe and accurate models

These comments concern the collection of training data. Whether a trained model itself processes personal data is a separate question, assessed case by case under EDPB Opinion 28/2024, and falls outside the scope of these Guidelines

Several passages treat the breadth of collection as a risk in itself. Para 8 links untargeted scraping to limited knowledge of the data collected, para 52 states that narrowing collection criteria "may be crucial" to meet the necessity condition, and paras 57-58 tie the level of interference to the number of websites scraped.

General-purpose models need broad and diverse data to understand language and context, to reduce bias and to produce accurate outputs. Removing personal data wholesale would prevent a model from learning how language refers to people and objects.[5] Much of a broad corpus is not personal data at all (code, scientific and factual text), so narrowing collection mainly reduces quality and representativeness rather than personal data processing. The AI Act's emphasis on relevant and representative data points in the same direction, and the two frameworks should be read coherently.

[4] EDPB-EDPS Joint Opinion 2/2026, recommending that proposed Article 9(5) GDPR emphasise safeguards implemented across the AI development lifecycle. See also *Legal analysis of the Digital Omnibus proposal amending Article 9 GDPR*, Morrison & Foerster for DOT Europe, paras 19 and 44-45.

[5] *Legal analysis of the Digital Omnibus proposal amending Article 9 GDPR*, Morrison & Foerster for DOT Europe, para 16: deleting all personal data from training sets would prevent a model from interpreting natural language in context, and current anonymisation techniques cannot yet preserve that context.

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

The preference for synthetic data (paras 37, 38 and 52) raises a similar issue. Synthetic data is a useful complement, but it cannot replace real-world data: it may not yet offer sufficient quality or diversity, and the models that generate it must themselves be trained on real data.

We recommend that the EDPB:

- recognise that broad crawling, combined with appropriate safeguards, is a standard and proportionate method for developing general-purpose models, and that the necessity test should focus on those safeguards rather than on breadth as such (paras 8 and 52. Safeguards such as source selection, exclusion of high-risk sources, filtering, deduplication, de-identification, security and access restrictions, commitments not to re-identify, rights facilitation, memorisation and extraction testing, and output controls should be given meaningful weight.
- present synthetic data as one option among others, not as a default preference (paras 37, 38 and 52). Indeed, if synthetic data or pseudonymisation are interpreted as default requirements, the guidance could undermine legitimate AI training where fully synthetic data or full anonymisation would materially degrade the purpose or model quality. The final Guidelines should clarify that synthetic data, anonymisation, and pseudonymisation controls are proportional tools, not mandatory defaults.
- give due weight to downstream controls under the minimisation principle (paras 36–38). Controllers cannot realistically screen the entire internet before collection, and para 68 itself recognises that it is not always possible to know what has been collected until after the data has been scraped. The final Guidelines should therefore clarify that controllers may comply through a documented combination of safeguards across the AI development lifecycle, selected as appropriate: source exclusions and collection filters where feasible, post-collection deletion or pseudonymisation and, where necessary, access, retention, training, testing and output controls to prevent unnecessary or sensitive data from being revealed or inferred.

4.2 Transparency should be meaningful and proportionate

DOT Europe supports transparency that helps individuals understand the processing and exercise their rights. Some recommendations, however, go beyond what is feasible or useful.

- **"Routinely" (para 27).** Article 14(5)(b) GDPR applies whenever its conditions are met, which the criteria in para 28 already test case by case. The suggestion that it should not be relied on "routinely" creates uncertainty and should be removed.
- **Transparency on source-level information (paras 30, 31 and 66(e)).** Article 14(2)(f) GDPR requires information on the source of the data, including whether it came from publicly accessible sources, not a URL-level index. Listing all domain names, URLs and collection dates in searchable form, or maintaining an updated list of scraped websites, is not feasible for corpora spanning billions of pages. It would also expose commercially sensitive information including trade secrets and security-relevant information, for example by facilitating data poisoning, extraction of membership-inference attacks. Nor would it help individuals: a list of thousands of pages that may or may not contain their data would increase confusion rather than transparency. Building on the fallback that para 30 already provides, the final Guidelines should confirm that information on categories of sources and

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

domain types, together with the crawler's characteristics mentioned in para 30, is sufficient where a fuller list would be disproportionate or create such risks, and that the searchable disclosure described in para 31 remains good practice only.

- **Consistency with the AI Act.** General-purpose AI providers already publish a summary of training content under Article 53(1)(d) AI Act, following the Commission's template.[6] The Guidelines should align with that summary rather than add a parallel, more detailed disclosure regime.
- **Publication of DPIAs (para 34).** Article 35 GDPR does not require publication, and DPIAs contain confidential and security-sensitive information. They should remain available to supervisory authorities on request. Confidentiality also supports candid assessments and the prior consultation of supervisory authorities under Article 36 GDPR, and avoids exposing vulnerabilities that have been assessed but not yet fully mitigated. Publishing a non-confidential summary could be mentioned as an optional good practice, provided it excludes information such as detailed sources list, filtering rules and security controls.
- **Example 2.** Excluding all web pages where individuals are directly identifiable would remove most news, encyclopaedic and public-interest content, including information about public figures. We suggest replacing it with targeted measures, such as the format-based filtering described in para 38 and the exclusion of high-risk source categories (for example, sources known to host leaked data or predominantly sensitive content).

4.3 Making the balancing test operable

The balancing test would benefit from clearer contours. Example 7 concludes that the test may generally be considered as met where a particular combination of safeguards is in place. The final Guidelines should add indicative examples of realistic safeguard combinations that would generally satisfy the balancing test, for example deleting or minimising personal data after collection, publishing category-level transparency, and excluding sources with significant personal or sensitive data. The benefits of the processing identified when assessing the legitimate interest (paras 43–50), including for users and society, should also be given due weight.

- **Training and deployment risks (para 54).** The risks described, such as de-anonymisation through web search tools, profiling and deep fakes, arise mainly when a model is used. Training extracts statistical patterns; its purpose is not to profile individuals. Para 61 itself provides that the likelihood of further consequences should be assessed taking into account the technical and organisational measures in place. The same applies to the suggestion that training creates a sense of surveillance: training does not monitor individuals or act on information about them. And memorisation is actively mitigated. Impact should therefore be assessed realistically, at the level of the individual data subject and in light of the safeguards applied. The Guidelines should distinguish training-stage from deployment-stage risks and recognise that output controls and usage policies mitigate the latter.

[6] Article 53(1)(d) Regulation (EU) 2024/1689 (AI Act) and the Commission's template for the public summary of training content for general-purpose AI models.

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

- **Credit for all effective safeguards (para 65).** Excluding measures the GDPR requires anyway from the balancing test risks overlooking safeguards that, in practice, significantly reduce impact. We invite the EDPB to clarify that the overall package of safeguards is assessed, including where controllers implement legal requirements in ways that go beyond the minimum.
- **Robots.txt and similar signals (paras 11, 37, 64 and Example 5).**We support the exclusion from collection websites that clearly oppose scraping through established, machine-readable signals such as robots.txt, applied at domain or URL level. These signals are scalable, widely used and can evolve through standardisation to address further needs. Respecting them should be recognised as an important safeguard in the balancing test and as relevant to reasonable expectations (para 64). However, the Guidelines should not require controllers to honour non-standardised formats, or to build new individual-level opt-out mechanisms or registries.
- **Opt-out lists (para 66(f)).** Asking individuals to provide identifying information so their data can be excluded would create new databases of personal data. It also sits uneasily with Article 11 GDPR, which the Guidelines recall in para 29. Existing channels are more privacy-friendly: the right to object via public notices, machine-readable opt-outs by website operators, and output-level controls. Where a data subject objects in relation to a trained model, output-level controls are also the proportionate way to give effect to the objection, much as search engines act at the point of access rather than on the underlying source (*Google Spain*, C-131/12).
- **Example 7.** The example sets a standard few models could meet in practice. It relies on content that data subjects have "manifestly made public", a concept from Article 9(2)(e) GDPR that the CJEU interprets strictly.[7] It also makes copyright compliance a factor in the GDPR balancing test, although copyright is governed by the DSM Directive and Article 53(1)(c) AI Act.[8] Whether copyright rules have been respected is a matter for the competent courts and, as regards text and data mining opt-outs, for the AI Office under the AI Act, rather than for the GDPR balancing test. We recommend removing the copyright element and replacing the reference to the content “manifestly made public” with a reference to publicly accessible online sources, as one relevant factor in the balancing test. .

4.4 Accuracy should be assessed against the purpose of training

AI companies already apply quality filtering and favour reputable sources. The recommendations in para 42 would, however, be difficult to apply to large-scale training:

- **Reliable sources.** Using official sources such as government registries as the benchmark would produce skewed, unrepresentative models. The notion of reliable sources should cover diverse, good-quality content, with low-quality content filtered out.
- **Timestamping and spot-checks.** Recording the scraping date of each item and spot-checking random samples for factual accuracy does not give meaningful assurance at this

[7]CJEU, C-252/21, *Meta Platforms v Bundeskartellamt*, 4 July 2023, ECLI:EU:C:2023:537, para 77.

[8]Article 4(3) Directive (EU) 2019/790 (DSM Directive); Article 53(1)(c) AI Act.

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

scale. Checking facts about private individuals would also require collecting more data about them, which runs counter to minimisation.

- **Focus on outputs.** Accuracy matters most where a model produces personal data as output, as para 40 recognises. For a general-purpose model, outputs depend on the user's prompt, the context supplied and how the model is deployed. The controller can check sources and remove obvious errors, but it cannot guarantee correct personal information in every output. The Guidelines should therefore focus on reasonable measures proportionate to the deployment context, such as output controls, user feedback channels and processes to correct reported inaccuracies, rather than implying that every output must be accurate.
- **Data mapping (para 37).** The EDPB should clarify that records of processing activities under Article 30 GDPR, satisfy the data mapping and inventory expectation.

4.5 Special categories of data

The application of *GC and Others* in paras 72–74 is the most important step forward in the draft Guidelines. Effective measures should mean genuine, documented and iterative effort proportionate to the state of the art, not a guarantee of zero leakage. The following clarifications would make it workable in practice and consistent with the Digital Omnibus negotiations.

- **Similarity with search engines (para 73(a)).** This condition should be assessed functionally. Like search engine operators, AI developers do not publish the data; they intervene after it has been made available online. Their responsibility arises from what the model may output in response to a prompt, much like a list of search results in response to a query.[9] The comparison should therefore turn on the structural features of the processing: automated, large-scale collection from publicly accessible sources, without individual review of items and without the ability to predetermine which categories of data will be encountered.
- **Information about public figures (paras 67 and 73(b)).** The Guidelines distinguish only between intentional processing and incidental or residual processing. Information about public figures, such as a politician's party affiliation or a head of state's publicly reported illness, belongs to the public domain and is needed for models to answer accurately. It is neither residual nor collected to profile individuals. We invite the EDPB to clarify, consistent with *GC and Others* and CNIL guidance, that such data may be processed where it relates to the person's public role. The controller would then verify on request whether a derogation applies, such as Article 9(2)(e) GDPR where the person has manifestly made the information public, or Article 9(2)(g) GDPR read with Article 11 of the Charter.[10] The same logic should apply to deletion requests under para 74(ii).
- **Reasonable steps before and after collection (para 74(i) and (ii)).** Filters that "prevent the collection" and deletion "immediately after the collection" read as absolute obligations. Guaranteeing that no special category data is collected would require pre-screening the

[9] Legal Assessment, paras 27–28.

[10] *GC and Others*, cited above; CNIL, "AI: respecting and facilitating the exercise of data subjects' rights" (example of a public figure's erasure request). See also DOT Europe Legal Assessment, paras 25, 28–30 and 33.

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

entire web, which is not possible with current technology (see 4.1). This is compounded by the fact that special category data is often contextual rather than structured, so pre-collection filters produce high false-positive rates. Systematically identifying such data also involves additional processing and inference, which can be more intrusive than the incidental presence of the data itself. We recommend referring to "reasonable steps" and deletion "as soon as reasonably practicable". The Guidelines should also distinguish deletion from pre-training datasets, which can be carried out within a reasonable time, from deletion after training, which would require costly retraining or unlearning.

- **A holistic assessment (para 74(iii) and (iv)).** Para 74 currently expects near-perfect results at every stage, from ingestion to outputs. Consistent with section 3, effective memorisation mitigation and output filtering should be able to compensate for the technical limits of pre-training filtering. This mirrors the lifecycle approach recommended by the EDPB and EDPS for proposed Article 9(5) GDPR.[11] Similarly, the expectation in para 74(iii) to provide assurance of state-of-the-art resistance to privacy attacks should be met by documented red-teaming, memorisation testing and extraction testing.
- **Monitoring (para 74(iv)).** "Constantly" monitoring outputs should be replaced by regular, risk-based monitoring, in line with para 76. Where a model is released with open weights, the developer cannot monitor downstream outputs. Para 74(iv) already places output monitoring on "the controller using the model as part of an AI system". The final Guidelines should confirm that this means the deployer, consistent with the "responsibilities, powers and capabilities" test.
- **Criminal offence data (Article 10 GDPR).** Paras 70–72 refer to Articles 9 and 10, but para 74 addresses only Article 9(1). As *GC and Others* covered both provisions, the Guidelines should confirm that the same framework applies to Article 10 data.[12]

4.6 Scope: user-directed retrieval and integrity and security crawling

The Guidelines focus on scraping for the training of generative AI. They should confirm that real-time, user-directed, inference-time retrieval (agentic browsing, tool use, retrieval-augmented generation) falls outside their scope. This includes targeted single-URL fetches initiated by the user for an immediate response, which are transient and not stored for training, analogous to a search engine answering a query or a user clicking a link. The bulk-scraping framework (exclusion of opposing websites, source-level transparency, pre-collection filtering) is not designed for such fetches, and the appropriate legal basis will depend on how the service is set up. Any subsequent reuse of such data for model improvement should be addressed separately, with requirements proportionate to the nature and scale of the reuse. The Guidelines should also confirm that crawling for integrity and security purposes falls outside their scope. Crawlers operated to detect malicious content, verify link safety and support anti-fraud investigation commonly disregard robots.txt precisely so that protection is not compromised. Para 50 already recognises fraud

[11] EDPB-EDPS Joint Opinion 2/2026; DOT Europe Legal Assessment, paras 38–39 and 44–45, and section X (alternative wording for Article 9(5) GDPR).

[12] Legal Assessment, paras 19 and 37, and section X (proposed extension of the derogation to Article 10 GDPR).

Avenue Marnix, 17
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu

detection and threat detection as legitimate interests. Many organisations also maintain a bright-line prohibition on reusing integrity- and security-crawled data for AI training, evaluation or fine-tuning, which demonstrates purpose limitation in practice. Such ring-fencing should be recognised as evidence of purpose limitation, so that crawling which bypasses robots.txt for these purposes is not treated as a negative factor when assessing an organisation's scraping for AI training.

5. Conclusion

DOT Europe supports the EDPB's objective and many elements of the draft Guidelines. To make them workable in practice, we invite the EDPB to focus on three priorities:

1. frame safeguards as a proportionate, lifecycle-based toolbox, assessed against a "reasonable steps" standard;
2. set transparency at the level of source categories, align with the AI Act training content summary;
3. make the *GC and Others* framework workable, including for information about public figures and for Article 10 data.

These changes would keep the Guidelines consistent with the AI Act and the ongoing Digital Omnibus negotiations, and give controllers and supervisory authorities a durable common reference. DOT Europe and its members remain available to share technical expertise and to continue this dialogue.

Avenue Marnix, *17*
1000 Brussels,
BELGIUM

TEL
+32 (0)472 26 83 02

MAIL
info@doteurope.eu

WEBSITE
www.doteurope.eu