



European AI Office Working Group meetings: Code of Practice for General-Purpose AI Template for summary of training data

AI Office
Working Group 1 - Copyright-related rules
17 January 2025

1. Template for summary of training content

Provisions in the AI Act

Article 53(1)d)

Providers of general-purpose AI models must draw up and make **publicly** available a **sufficiently detailed summary** about the content used for training of the GPAI model, according to a **template** provided by the AI Office.

Recital 107

- The summary must be **comprehensive in its scope** rather than technically detailed
- **To facilitate parties with legitimate interests** (including rightholders) **to enforce their rights**, while taking due account of **providers' trade secrets**
- AI Office's **template** should be **simple, effective** and allow the summary in a **narrative form**

1. Template for summary of training content

Process and timeline

Template for summary of training data

- To be **adopted by the Commission** together with explanatory guidelines on general-purpose AI rules
- **Complementary to the Code of Practice** (the process is linked)
- The template to **set the structure and minimum content** required for all GPAI models (signatories to the Code or not)
- The Code of Practice can include further voluntary commitments to ensure '**adequacy of detail**' of the summary (Art. 56(1)b))
- Closely **interlinked with the policy on copyright** (Art. 53(1)c)AI Act)

Consultation:

Broad stakeholder consultation on GPAI ~ 430 responses (30 July-18 Sept. 2024)

- Informed first draft by Chairs of the Code of Practice
- Informed the AI Office's drafting of the template for the summary of training data

Next steps:

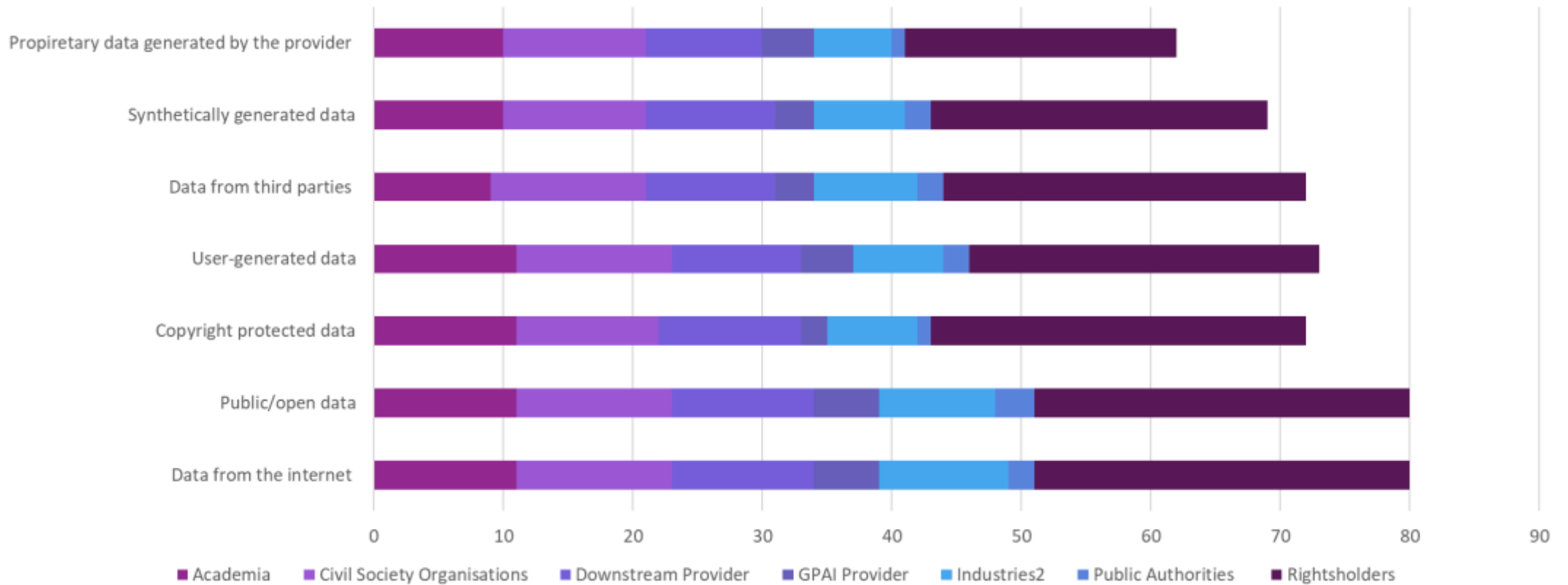
- 17 January:** Presentation of AI Office's draft to Code of Practice WG 1 Transparency and copyright,
- By 31 January:** Written feedback by stakeholders from WG1
- End of February:** Presentation of the template and draft GPAI guidelines to Code of Practice WGs
- Q2 2025:** Adoption of template and GPAI guidelines
- 2 August:** Entry into application of general-purpose AI rules

2. Multistakeholder Consultation

Categories of information sources

1

Categories of information sources that should be presented in the summary?

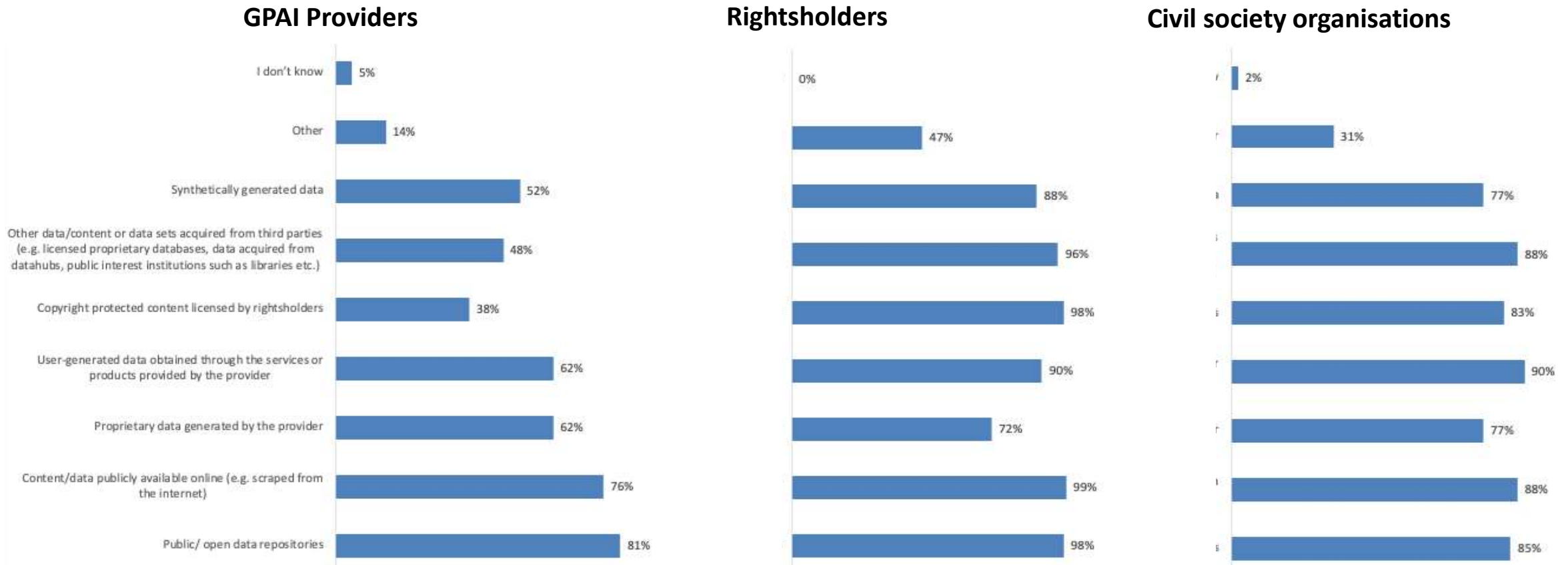


2. Multistakeholder Consultation

Categories of information sources

1

Categories of information sources that should be presented in the summary?

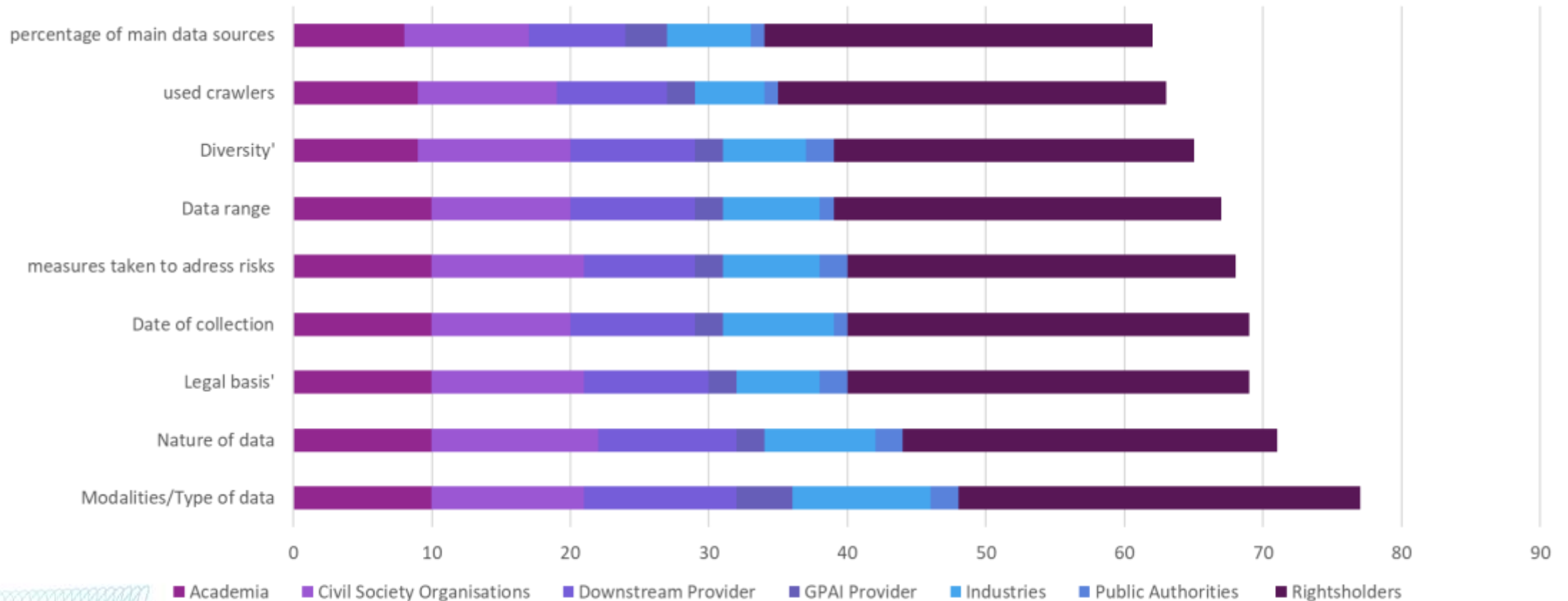


2. Multistakeholder Consultation

Information about the data sources

2

Should the summary include one or more of the following characteristics/information about the data used for the training?



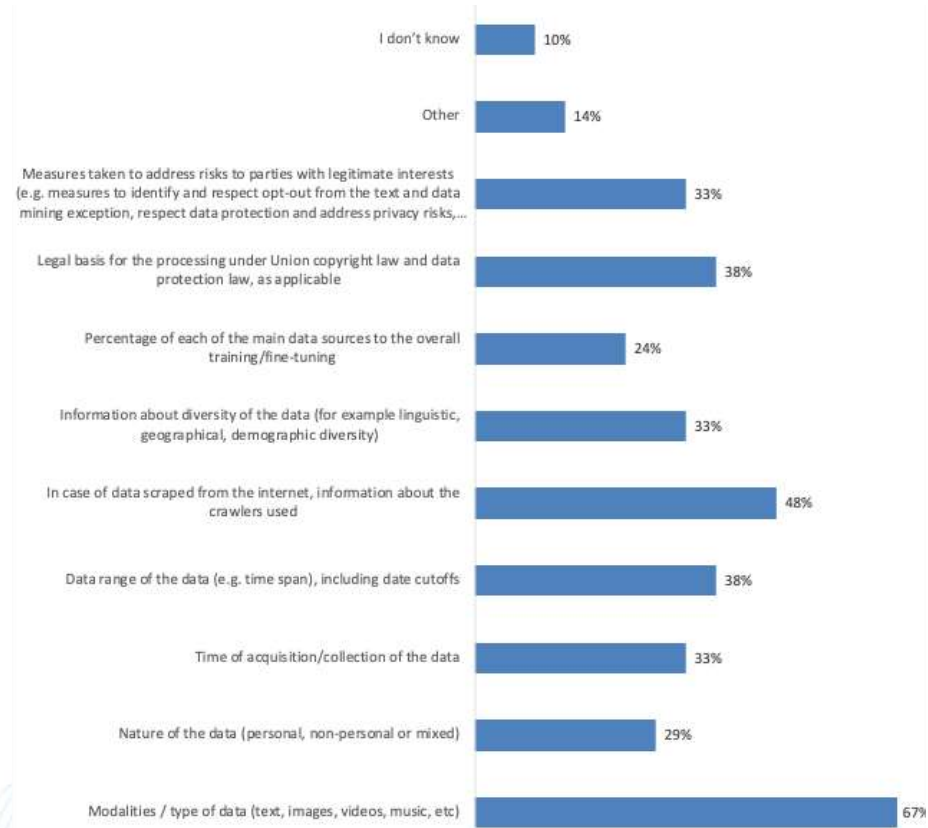
2. Multistakeholder Consultation

Information about the data sources

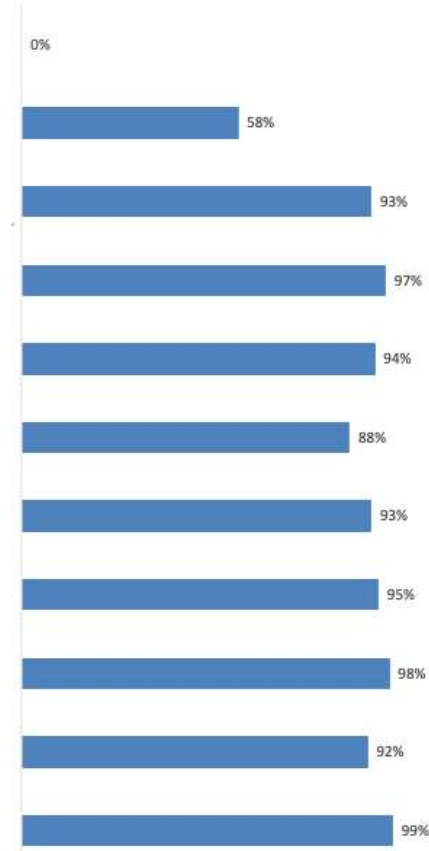
2

Should the summary include one or more of the following characteristics/information about the data used for the training?

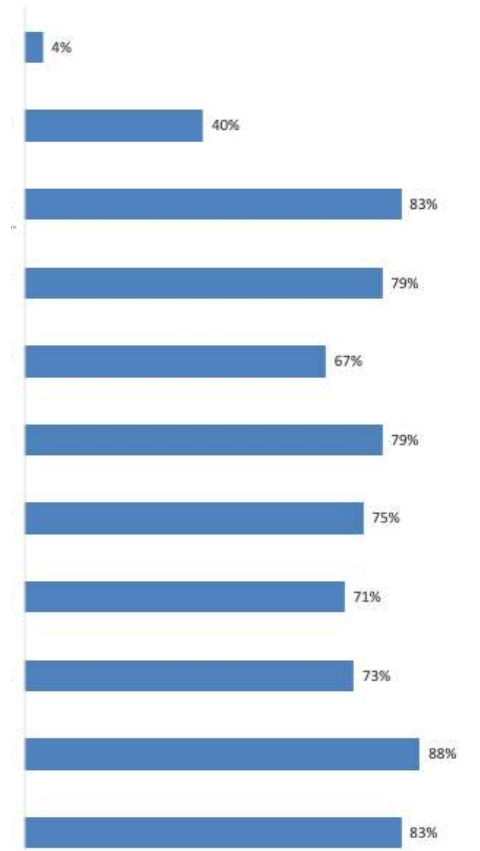
GPAI providers



Rightsholders



Civil society organisations



2. Multistakeholder Consultation

Main Results: Balance needed

3

Open Question: Further recommendations and balance with trade secrets?

Trade Secrets

Industry/Providers

- Too detailed disclosure can lead to security **vulnerabilities** or **prejudice to AI providers'** competitive advantage
- Trade secrets and confidential information should be **interpreted broadly**
- **Against** disclosure of:
 - Algorithms and model architecture
 - Concrete works/URLs
 - Specific data treatment processes
 - Precise composition/percentage
 - Legal advice

Transparency

Rightsholders/NGOs

- **Work-by-work disclosure of all sources of data** to train and test the GPAI
- Trade secrets should **not be used to prevent the disclosure** of the information which is essential
- For synthetic data, the summary should also include information on the **model** and "**data-cleaning standards**"
- **Regulatory oversight** to ensure compliance
- High public transparency of all data sources and lawfulness
- Mitigating measures to address risks to fundamental rights

3. AI Office's approach to the template

Key principles and structure

- **Objective:** to facilitate parties with legitimate interests to exercise their rights under Union law (e.g. rightholders, data subjects)
- **Completeness:** cover all sources of content from pre-training to fine-tuning
- **Effective:** with sufficient details to achieve its objective
- **Simple:** Non-technically detailed, understandable, no legal expertise needed
- **Balance with trade secrets:** due account taken of the need to protect trade secrets, disclose the ingredients but not the secret recipe
- **Proportionate:**
 - Summary narrative form
 - For fine-tuned models: only the additional data used
 - SMEs proportionate burden
 - Up to date: for small updates of the model not more often than 6 months

Proposed sections:

1. General Information

2. List of Data Sources

3. Relevant Data Processing
Aspects

3. AI Office's approach to the template

Section 1 General information



1.1 Model and provider identification

- Provider's name and contact
- Authorized representative
- Model identifier
- Base model(s)

1.2. Date of placement on the market and knowledge cut off date

1.3. Overall training data size, modalities and characteristics

Modalities	Overall size
<input type="checkbox"/> Text	Number of tokens or bytes
<input type="checkbox"/> Image	Number images (or pairs with other media)
<input type="checkbox"/> Video	Number of minutes (or pairs with other media)
<input type="checkbox"/> Audio	Number of minutes (or pairs with other media)
<input type="checkbox"/> Other	_____[please specify]

- Description of the **linguistic, regional, demographic and other relevant characteristics** of the overall training data:

Text	Image	Video	Audio
<input type="checkbox"/> Fictional texts, literature <input type="checkbox"/> Scientific and educative texts <input type="checkbox"/> News, journalism and opinions <input type="checkbox"/> Legal and official documents <input type="checkbox"/> Social communication (e.g.messages) <input type="checkbox"/> Promotion, advertising, product and service reviews <input type="checkbox"/> Other text	<input type="checkbox"/> Photography <input type="checkbox"/> Paintings & fine-arts <input type="checkbox"/> Infographics <input type="checkbox"/> Illustration & graphic design <input type="checkbox"/> Social / personal images Special <input type="checkbox"/> Source code <input type="checkbox"/> Structured data (e.g. calendar, maps) <input type="checkbox"/> Other, describe:	<input type="checkbox"/> Movies, shows, performances <input type="checkbox"/> Animated video content <input type="checkbox"/> Video game & immersive footage (e.g. 3D) <input type="checkbox"/> Documentaries <input type="checkbox"/> Video news and journalism <input type="checkbox"/> User content, short videos <input type="checkbox"/> Other video content (e.g. experimental art,	<input type="checkbox"/> Music <input type="checkbox"/> Narrative and fiction (e.g. audiobooks) <input type="checkbox"/> Non-fiction educative audio content <input type="checkbox"/> Radio shows and podcasts <input type="checkbox"/> Social communication (phone calls, voice messages) <input type="checkbox"/> Other (e.g. sounds and ambient)

3. AI Office's approach to the template

Section 2 List of data sources



2. List of Data Sources

2.1. Publicly accessible datasets:

- Overall size per modality and number of all datasets (number of synthetic datasets)
- List of 'main/large' datasets (above 5% of the overall data in this category) with unique identification, links + period of collection

2.2. Private non-publicly accessible datasets of third parties:

- Data licensed by rightholders or their representatives: Overall size per modality
- Datasets acquired from other third parties: Overall size per modality and number of datasets (number of synthetic datasets)
- List of 'main/large' private data sets acquired from other third parties (above 5% of the overall data in this category), unique identifiers and links (if available) and narrative description + period of collection

2.3. Data crawled and scraped from online sources:

- Overall size per modality, period of scraping
- Identification of crawlers, their purpose and behaviour;
- Explanation what content has been targeted;
- List of top 10 % of all internet domain names per type of data modality (e.g., text, image).
- For SMEs top 5 % or 1 000 internet domain names regardless of data modalities, whichever is lower, unless the model is with systemic risks.

2.4. User-sourced data (collected by provider incl. prompts):

- Overall size per modality
- List of providers' services/products

2.5. Self-sourced synthetic data(sets):

- Overall size per modality
- Name of AI model

2.6. Data acquired by the provider through other means:

- Overall size per modality
- Means of acquisition

3. AI Office's approach to the template

Section 3 Relevant data processing



3. Other Relevant Data Processing Aspects

2.1. Respect of copyright and related rights

- Measures implemented to respect reservations of rights from the text and data-mining exception under Art.4(3) DSM Directive **during** data collection incl. specification of the opt-out protocols and solutions honoured by the provider
- Measures implemented **after** data collection is completed to identify and remove content for which rights have been reserved by the rightsholders

2.2. Removal of unwanted content

- Describe content deemed unwanted by the provider as part of the training data
- List the measures taken to avoid and/or remove such content (such as blacklists, keywords, and model-based classifiers)
- Measures applied by the curators of listed datasets may be mentioned, but do not need to be listed exhaustively