1 Questionnaire Design

To set up our questionnaire, we decided to use a questionnaire designed and tested by experts in the field. After a review of the literature, we identified several relevant questionnaires [6, 1, 9, 5, 3]. In order to choose the most appropriate, in terms of completeness and coverage of GDPR, we carried out two types of evaluation, a qualitative and a quantitative one. By combining these two methods, we can make an informed decision on which questionnaire to use for our study. These evaluations are presented below.

Qualitative Assessment We reviewed each of the questionnaires and found that the questions were grouped into categories. However, categories and questions differ from one questionnaire to another. Therefore, we cannot consider these categories as direct criteria for comparison. So, we did a second reading and compared which aspects of GDPR were covered by the questionnaires. We then analyzed the questions one by one[1]. Given the large number of aspects, in order to reduce the complexity of the comparison table, we grouped these aspects into the following classes:

- Lawfulness: consent, consent and decisions for minors, legal basis, record of processing activities (register).
- User rights: to be informed, withdrawal of consent, access, portability, erasure, rectification, restriction, objection, automated decision making, and profiling.
- Accountability and governance: purpose, accuracy, minimization, data retention, third parties, DPO, DPIA, representative.
- Security and breaches.
- Transfer (outside the EU).

These groups form the comparison criteria for our qualitative analysis. The number of questions criterion gives an indication of the level of detail of the questionnaire. Table 2 shows the results expressed as normalized scores. The questionnaire from [6] has the highest GDPR coverage score, with 22 covered aspects out of 24.

Quantitative Assessment In order to confirm the results obtained in the qualitative assessment, we conducted an automated NLP-based evaluation[2]. We first compare each questionnaire to a set of key terms from GDPR. To do this, we used the French-English data protection glossary of the official website of the CNIL [2], which contains 430 key terms, as well as the titles of the 99 articles of the regulation [4]. We consider that these two sources are strongly representative and include the key terms of GDPR. In order to be exhaustive, we then conducted a second comparative study to compare the questionnaires with the full text of the GDPR.
The adopted approach consists in comparing (by measuring the similarity between) the vectorized representations of each document (the key terms, the full text of GDPR and the questionnaires). These vectors are obtained by common document embedding techniques (*see* below). The followed process is presented in Figure 1.

- Data Collection. The goal of this step is to record each questionnaire and GDPR terms / regulation text in separate text files. The questionnaires [6, 1, 5, 3], the French-English data protection glossary and the titles of the GDPR articles[3] are available online. For extracting the questionnaire [9], we accessed the GitHub repository of the application[4]. This data collection was done manually.

[1]Details of all the questions and the keywords covered by each are available at `https://github.com/gitPriam/PRIAM_Metamodel/`

[2]The code implemented for this evaluation is available here: `https://github.com/gitPriam/PRIAM_Metamodel/`

[3]https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[4]https://github.com/ElsevierSoftwareX/SOFTX-D-20-00011

Table 1: Qualitative assessment of questionnaires.

Questionnaire / Criteria	Number of questions / Criteria	Lawfulness				User rights									Accountability & Governance									Security & Breaches		Transfer	Total /24
		Consent	Consent and decisions for minors	Legal basis	Record of processing activities	To be informed	Withdrawal of consent	Access	Portability	Erasure	Rectification	Restriction	Objection	Automated decision making and profiling	Purpose	Accuracy	Minimization	Data retention	Third parties	DPO	DPIA	Representative	Security	Breaches	Transfer outside EU	
[6]	54	×	×	×		×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×	×	22
[1]	31	×	×	×	×	×		×	×	×	×	×	×	×		×			×	×	×		×	×	×	19
[9]	31	×				×		×		×	×				×	×	×	×					×			10
[5]	19				×	×		×	×	×	×	×	×	×					×	×	×	×	×	×		15
[3]	32	×		×	×	×	×	×	×	×	×	×	×	×	×		×		×		×	×	×	×	×	20

- Pre-processing. This step consists in transforming each collected text to a more suitable form for machine processing. This step is very important because the quality of the data can impact the results. The pre-processing step is a process in itself composed of several operations:
  - Lowercase the text
  - Remove punctuation
  - Tokenize the text
  - Remove stop-words
  - Remove duplicate words
  - Remove numbers
  - Initialize the lemmatizer
  - Remove whitespace tokens
- Document vectorization. Once the documents have been pre-processed, they are converted into numerical vectors that embed abstract representation of their contents (projections into multi-dimensional vector spaces). Similar contents are expected to produce similar vectors. Calculating similarities between the vectors embedding documents is thus used as a similarity measure between the contents of the corresponding documents. We use this approach to evaluate similarities between the questionnaires and the "GDPR key terms" or the "GDPR full-text" documents. Two techniques were tested:
  - **TF-IDF** measures the importance of a word in documents (based on word frequency). Because of the number of documents used to compute this metric, TF-IDF did not produce relevant results. This technique was thus set aside to rather focus on the following one.
  - **Word Embedding** uses neural networks to train a model on hundreds or thousands of documents. The goal is to assign to each word a numerical vector representing its semantics as contextual relationships with other words. This technique has been adapted to embed directly the semantic of sentences, paragraphs, and whole documents. However, the models have limitations with respect to their input layers, such as the size of the documents that can be handled. A solution is to truncate the documents, which leads to the issue of selecting the parts that best represent their contents. Another solution is to use the averaging of the embedding vectors of the words of a document as an embedding vector for the document. We have used this second solution.

Results obtained with TF-IDF were compared to those of two pre-trained word embedding models:

- Word2vec [8]: a model trained on 3 million words from Google News data. It is provided by Google and is a 300-dimension model.
- FastText [7]: a model trained by Facebook's AI Research Lab (FAIR) on Common Crawl and Wikipedia. It is also a 300-dimension model.

The result of this step is the vector representations (embeddings) of each document (the average of the word vectors).

- Similarity Calculation. This step consists in measuring the similarity between the questionnaires, the "GDPR full text" and the "GDPR key terms" documents once their mean word vectors have been calculated. The well-known cosine similarity [10] was used as a metric.

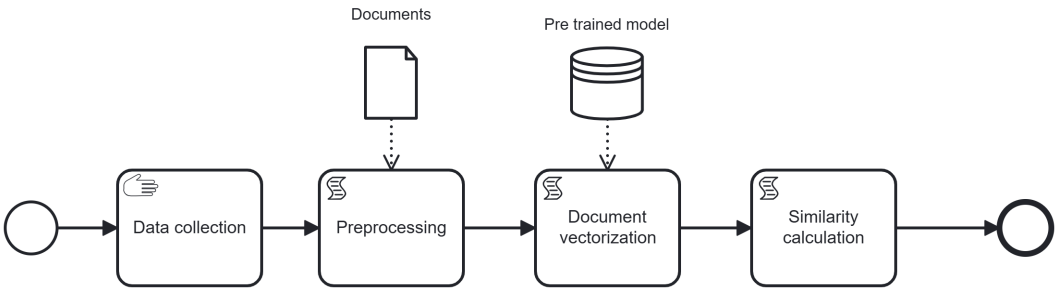

Figure 1: NLP-based similarity calculation process.

Table 2 presents the results obtained when using the two word embedding models (Google's word2vec and fastText's Common cowl) and two lemmatization techniques (WordNet and Spacy lib). The two word embeddings are presented in columns, while the two lemmatization techniques are presented in rows. The results of the similarity measure per questionnaire, for each combination of word embedding and lemmatization, are presented in the cells.

Table 2: Comparison of the questionnaires with the key terms and full text of GDPR.

Type of lemma	**Questionnaire**	**Full text**		**Keys terms**	
		Word2vec	**FastText**	**Word2vec**	**FastText**
WordNet (NLTK lib)	[6]	**0.917**	**0.924**	**0.947**	**0.927**
	[1]	0.887	0.866	0.921	0.866
	[9]	0.886	0.838	0.886	0.839
	[5]	0.887	0.876	0.884	0.865
	[3]	0.902	0.898	0.925	0.884
Spacy lib	[6]	**0.924**	**0.933**	**0.939**	**0.897**
	[1]	0.902	0.881	0.915	0.837
	[9]	0.879	0.85	0.885	0.814
	[5]	0.903	0.892	0.888	0.854
	[3]	0.914	0.912	0.92	0.859

Figures in bold represent the highest similarity values. The results in Table 2 show that questionnaire [6] has the highest similarity to the "GDPR key terms" and "GDPR full text" document followed by [3], regardless

of the word embedding or lemmatization used method[5]. The two word embedding models produce robust, identical results, consistent with the qualitative assessment. Consequently, we selected questionnaire from https://www.dataprotection.ie/ [6].

References

[1] Controllers checklist—ico [online]. available at. https://ico.org.uk/for-organisations/sme-web-hub/checklists/data-protection-self-assessment/controllers-checklist/. (Accessed on 06/04/2025).

[2] English-french glossary on data protection, CNIL. https://www.cnil.fr/en/english-french-glossary-data-protection. (Accessed on 06/04/2025).

[3] EU GDPR readiness assessment tool [online]. available at. https://advisera.com/tools/eu-gdpr-readiness-assessment-tool/. (Accessed on 06/04/2025).

[4] GDPR archives - gdpr.eu. https://gdpr.eu/tag/gdpr/. (Accessed on 06/04/2025).

[5] GDPR checklist for data controllers [online]. available at. https://gdpr.eu/checklist/. (Accessed on 06/04/2025).

[6] Preparing your organisation for the general data protection regulation - your readiness checklist [online]. available at. https://www.dataprotection.ie/sites/default/files/uploads/2019-04/A-Guide-to-help-SMEs-Prepare-for-the-GDPR.pdf. (Accessed on 06/04/2025).

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] Fábio Pereira, Paul Crocker, and Valderi RQ Leithardt. PADRES: Tool for Privacy, Data REgulation and Security. *SoftwareX*, 17:100895, 2022.

[10] Jiapeng Wang and Yihong Dong. Measurement of text similarity: a survey. *Information*, 11(9):421, 2020.

[5]Note that tests using different stemmers (Porter and Snowball) instead of lemmers were performed. The obtained results are not presented because their analysis has shown that a considerable number of words do not have a vector representation in the two tested word embeddings (on average, 145 words for FastText and 372 words for google-news).