## SUPPORT POOL OF EXPERTS PROGRAMME

# Al Possible Risks & Mitigations **Optical Character Recognition** by Isabel BARBERÁ



As part of the SPE programme, the EDPB may commission contractors to provide reports and tools on specific topics.

The views expressed in the deliverables are those of their authors and they do not necessarily reflect the official position of the EDPB. The EDPB does not guarantee the accuracy of the information included in the deliverables. Neither the EDPB nor any person acting on the EDPB's behalf may be held responsible for any use that may be made of the information contained in the deliverables.

Some excerpts may be redacted or removed from the deliverables as their publication would undermine the protection of legitimate interests, including, inter alia, the privacy and integrity of an individual regarding the protection of personal data in accordance with Regulation (EU) 2018/1725 and/or the commercial interests of a natural or legal person.

## Table of Contents

| 1. | Background   | 4  |
|----|--|----|
| 2. | Data protection and privacy risk identification  | 12 |
|    | Definition of the criteria to consider when identifying risks and their categorization | 12 |
|    | Presentation of examples of risks specific to OCR                                      | 13 |
| 3. | Data protection and privacy risk assessment  | 18 |
|    | Criteria to establish the likelihood of OCR risks. How to assess likelihood.           | 18 |
|    | Criteria to establish the severity of OCR risks. How to assess severity.               | 19 |
|    | Examples of OCR specific risks assessments   | 19 |
| 4. | Data protection and privacy risk treatment   | 23 |
|    | Risk treatment criteria  | 23 |
|    | Presentation of mitigation measure examples/risk treatment options                     | 24 |
|    | Residual risk acceptance   | 27 |
| Re | eference to specific technologies, tools, methodologies, processes or strategies       | 30 |

**Disclaimer by the Author**: the examples and mentions of companies in this report are illustrative and do not imply that the author considers them the only or the best choice. The technology analysis presented in this report is based on the state of the art of the technology in August 2023.

#### Document submitted in September 2023

## 1. Background

## Description of the task, main technologies used and references to some openly accessible examples.

OCR stands for Optical Character Recognition, and it is a technology used to convert images or scanned documents containing text into machine-readable text. OCR technology enables the extraction of text from both physical paper documents and digital sources.

## How do data extraction technologies like OCR work?

OCR techniques use different approaches such as rule-based methods and pattern matching algorithms to identify characters and convert them into machine-readable text.

The process of extracting data using OCR typically involves three primary tasks: detection, localization, and segmentation. Each of these stages can employ various algorithms.

- In the *detection* and localization stages, algorithms are employed to identify and locate text within a given frame or image.
- Localization algorithms analyze frames to determine the bounding regions surrounding the text, effectively pinpointing its location. These algorithms work together to identify and delineate text regions.
- The segmentation task involves converting the localized text into a binary format that is suitable for OCR processing. Segmentation algorithms apply techniques to transform the text into a format where characters are clearly distinguished from the background, improving the accuracy of character recognition.

In the *binarization*<sup>1</sup> process, characters are identified by recognizing dark areas as text and light areas as the background. The dark areas undergo processing to identify alphabetic letters or numeric digits. Algorithms for pattern recognition and feature extraction are then used to identify and analyze these characters.

*Pattern recognition* involves isolating a character image, referred to as a glyph<sup>2</sup>, and comparing it with a stored glyph that shares a similar font and scale. For successful pattern recognition, the stored glyph must closely match the input glyph in terms of font and scale. This approach is more effective when working with scanned document images that have been typed using a known font.

*Feature extraction* involves breaking down or decomposing glyphs into various characteristics, including lines, closed loops, line direction, and line intersections. These features are then used to identify the most suitable match among the stored glyphs.

In addition to character recognition, an OCR program examines the structure of a document image by segmenting it into elements like text blocks, tables, or images. After isolating the characters, the program compares them with a collection of pattern images. It then processes potential matches and presents the recognized text as the output.

<sup>&</sup>lt;sup>1</sup> Binarization is the step that is performed prior to performing OCR. The aim of binarization is to separate foreground text from the background of a document.

<sup>&</sup>lt;sup>2</sup> In typography, a glyph is "the specific shape, design, or representation of a character". It is a particular graphical representation, in a particular typeface, of an element of written language. Source: Wikipedia



Image source: Microsoft OCR Vision Studio. Example of segmentation and pattern recognition.

OCR systems often work with an established system of templates; this means that documents need to have the same basic page structure or the same relative positioning of elements within the document as the template. Because the system relies on defined templates, using OCR with documents that have a different structure will result in a lower accuracy. Currently OCR systems based on templates are available on a broad variety of languages and they provide and extensive catalogue of different templates.

Nowadays, modern OCR systems also incorporate Machine Learning (ML) algorithms, particularly those based on Deep Learning, to improve the accuracy of character recognition. This type of deep learning models can support documents that have similar information, but different page structures. This is called Intelligent Document Processing (IDP) and uses OCR as its foundational technology to additionally extract structure, relationships, key-values, entities, and other document insights. OCR combined with Deep Learning supports structured, semi-structured, and unstructured documents for data extraction.

Often these OCR systems are offered as Software as a Service (SaaS) solution offering the possibility to use pre-trained models or to train your own model with your own dataset<sup>3</sup>.

Most vendors offer OCR systems as a cloud solution via a system of APIs<sup>4</sup>, what seems to be the preferred option for most customers because of their ease of integration and fast productivity. Though some providers<sup>5</sup> of this technology offer a general system where the models are shared by the customers, there are also some that offer the possibility to have a custom model that the customer can train and delete when necessary<sup>6</sup>.

Some vendors also offer the possibility for customers to host the models on-premises making the OCR capabilities available in the customer's own local IT infrastructure. This can be a good alternative to comply with strict security and data governance requirements.

It is also possible to develop and implement your own OCR solution in-house. There are different OCR libraries and frameworks available such as Tesseract, OpenCV, Easyocr, Keras\_ocr and the FineReader engine from ABBYY.

<sup>&</sup>lt;sup>3</sup> Microsoft and ABBYY are examples of SaaS OCR solutions offering the two possibilities: https://www.abbyy.com/vantage/ocr-skill/features/

<sup>&</sup>lt;sup>4</sup> An Application Programming Interface is a way for two or more computer programs to communicate with each other (source: Wikipedia).

<sup>&</sup>lt;sup>5</sup> In 2023, some of the most known OCR solutions providers were ABBYY, Kofax and Microsoft.

<sup>&</sup>lt;sup>6</sup> https://learn.microsoft.com/en-us/azure/applied-ai-services/form-recognizer/overview?view=form-recog-3.0.0

| OCR SaaS solution hosted in      | OCR Third party solution hosted | OCR self-developed, hosted on |
|----------------------------------|---------------------------------|-------------------------------|
| cloud                            | on premises                     | premises                      |
| - Ready to use models that are   | Models trained by vendor or     | Models trained by user        |
| trained by the vendor            | customer                        |                               |
| - Possibility to create your own |                                 |                               |
| self-trained models              |                                 |                               |

## Data flow in an OCR solution:

OCR solutions are mostly used to digitize documents that are originally in paper format. Some OCR solutions can also be used for the extraction of data in documents that are already available in digital format. In both cases, the document that we want to digitize and analyze will be considered our **Input Data** and the results of the data extraction process will be the **Output Data**.

In the following examples we show three possible scenarios:

- 1. A customer uses an OCR third party solution hosted in the cloud
- 2. A customer uses an OCR third party solution hosted on premises
- 3. A user develops an own OCR system

#### 1. Example of data flow diagram when using third party OCR systems hosted in the cloud



**Step 1:** The input data are transmitted via an API from the customer's location where the OCR system is located to the vendor's location in the cloud where the data extraction process will take place.

**Step 2:** The input (and output) data can be temporarily stored locally at the vendor's location in the cloud. The most common storage options are the following:

#### Al Possible Risks & Mitigations - Optical Character Recognition (OCR)

1. The data could be stored in a buffer only during the execution of the data extraction process. The vendor does not retain any data once it has sent the output to the customer.

The data could also be temporarily stored in cache<sup>7</sup> to be reused by other immediate processes.
 The data retention period is variable and depends on the cache memory capacity and configuration.
 Another possible scenario is the storage of data in a persistent<sup>8</sup> storage layer such as a database or

a cloud storage. This could be done for the analysis or processing of the data at a later stage.

The longer the data is stored in a system the higher the risk of a data breach, unlawful repurpose or an infringement of the data storage limitation principle. In this specific case, number 1 (buffer) is the option with less risks since data is stored only during the process in memory. In option 2 (cache) though also with a low risk, data is usually stored for a longer period than in buffer and this can happen outside the process and even on a different location<sup>9</sup>. Option number 3 (storage location like a file or a database for instance) is the one with the highest risks since the storage can take place for a longer period of time.

**Step 3:** Once the data extraction process has finalized, the output data are sent back via an API to the customer.

In some cases, the input and/or the output data could be used by the vendor to retrain and fine-tune the OCR model. Though this is usually done after informing the customers and obtaining their consent, it is important to verify it with the vendor.



#### 2. Example of data flow diagram when using third party OCR systems hosted on premises

**Step 1:** All data transfers and data extraction process take place internally at the customer's premises on their own servers within their data centers.

<sup>&</sup>lt;sup>7</sup> Caching is usually implemented at a software level to reduce the computational overhead of reprocessing the same text or data and improve overall performance.

<sup>&</sup>lt;sup>8</sup> Persistent storage means that the data remains intact even when the system is powered off or restarted.

<sup>&</sup>lt;sup>9</sup> The location and method of caching can depend on the specific requirements of the OCR system and the architecture of the application. For instance, caching could happen in a different location when using a distributed microservices architecture or a cloud-based caching.

**Step 2:** The input and output data can be temporarily stored locally. The data could be stored in a buffer only during the execution of the data extraction process or could be temporarily stored in cache to be reused by other immediate processes.

The input and/or output data could also be stored in a storage location at the customer's premises. This could be done for analysis or processing of the data at a later stage or for auditing purposes.

Step 3: Once the OCR process has finalized, the output data is produced.

The input and/or output data could also be used to retrain and fine-tune the OCR model that is also stored at the user's premises.

A self-hosted OCR system from a third party provider can be set up in different ways depending on the architecture and design choices. The specific details can vary if the choice is a completely on premises set up or a hybrid one<sup>10</sup>in which some of the processes are still hosted at the vendor side. For instance, some of the steps in the data extraction process phase could happen outside the customers premises (see image below).



It's important to review the documentation and architecture of the third party OCR system to understand its data flows and whether there are data transfers to the vendor or other third-parties. Additionally, it is also important to assess the system's compatibility with the user's infrastructure, the security as well as any potential required change in network or firewall configurations.

<sup>&</sup>lt;sup>10</sup> Different steps from the data extraction phase could take place in the cloud. This could be a decision due to various reasons such as resource-intensive tasks, redundancy, expertise, etc.





The data flow in this scenario is similar to the one of example 2. In this use case, there are no data transfers to any OCR vendor, and all the processes are executed on premises. All processes take place internally.

#### Performance measure of OCR systems: Accuracy and Confidence Scores

The performance of OCR systems can be measured with different metrics being accuracy and confidence scores the two most commonly used.

The **accuracy** of an OCR system measures the percentage of correctly identified characters or words with respect to the total number of character or words. It is typically evaluated by comparing its output to the 'ground truth'<sup>11</sup> and calculating the proportion of characters or words that were correctly identified. Any discrepancies between the OCR output and the ground truth are considered errors. Higher accuracy rates indicate better performance. The accuracy value range is usually represented as a percentage between 0% (low) and 100% (high). For printed documents with clear and legible text, accurate OCR results in the range of 95% to 99% are commonly achievable. However, it's important to note that the accuracy can vary depending on the specific document types, languages, and OCR software being used.

**Confidence score** is a measure that provides an indication of the level of certainty of correctness associated with the extracted data. It is represented with a number between 0 and 1 and a high confidence score would mean that the OCR system believes its recognition of a particular text is likely to be correct. The confidence scores are typically determined by the OCR software or algorithm itself though in some OCR systems<sup>12</sup>users can set confidence score thresholds to filter out characters or results that fall below or above a certain level of confidence. This threshold can be set based on the desired level of accuracy and tolerance for errors. The specific method for calculating confidence scores may vary depending on the OCR system and the underlying algorithms used.

<sup>&</sup>lt;sup>11</sup> The accurately known text that the images or documents being processed by the OCR system are supposed to represent. The term "ground truth" is used in machine learning to refer to the actual values or outcomes that a model's predictions are compared against during training and evaluation. <sup>12</sup> https://pyimagesearch.com/2020/05/25/tesseract-ocr-text-localization-and-detection/

Confidence scores can be calculated in different ways, but the score is often based on the clarity of the image, character distinctness, context of surrounding text and the similarity of the identified characters to the patterns the system has been trained on.

While confidence scores at the character level are prevalent, OCR systems may also provide confidence scores at other levels such as word, line, or block.

The availability and presentation of confidence scores may vary among different OCR systems and software implementations. Some systems may only provide character-level scores, while others may offer a combination of character-level and higher-level scores.

While a high confidence score like 0.95 would suggests the OCR system believes the output is correct, it does not guarantee that the output is actually accurate.

If an OCR system persistently misread a particular character due to errors in its training data, it could still assign a high confidence score to this incorrect interpretation. Similarly, an accurate result could receive a low confidence score if the system finds the recognition challenging due to factors like image quality or unconventional font style. Hence, it's critical to consider these factors when interpreting confidence scores and accuracy in OCR systems.

Issues that can affect accuracy of the output:

- The accuracy of the models can be affected by variations in the structure of the documents. The accuracy scores can be inconsistent when the analyzed documents differ from documents used for training the model.
- The accuracy of the output is determined by the conditions and the quality of the input images.
   For instance, the system can be susceptible to variations in the position of a document (landscape or portrait), or to changes in fonts and formatting.
- OCR can introduce errors, such as incorrectly recognizing a character as another one. For example, the OCR could recognize "N" and change it to "E." This is common in texts with non-English characters. OCR might mistake a lowercase "l" for a "1", or a "b" for an "8". This can cause problems if the text is used for critical purposes, as could be the case with some legal documents.
- Punctuation marks cannot be always read by OCR because they are too small or non-contiguous, or because they're upside down and backwards.
- OCR may not be able to recognize text correctly if the text is in a language not supported by the OCR engine. It is important to verify that your language is supported. An OCR system might have difficulties recognizing properly right to left languages.

## Common uses of OCR technologies:

Currently, data extraction techniques like OCR are being used in different use cases. Here are examples of some of them:

- Accounts payable<sup>13</sup>: to speed up invoice data entry. OCR technology can be used to automate the
  process of entering invoice data into a system. Instead of manually keying in data from paper or
  digital invoices, the OCR software extracts key information such as supplier names, invoice dates,
  amounts, and invoice numbers. This helps to reduce data entry errors and speeds up the accounts
  payable process.
- Banking<sup>14</sup>: to extract and digitize information making data easier to search, store, and manage. In identity verification, OCR is used to read data from identity cards, passports, or driving licenses quickly and accurately. By extracting information from identity documents, OCR facilitates the verification of customer details and speeds up the account opening procedures.
- Digitizing<sup>15</sup> and/or archiving of paper documentation, converting printed paper documents into machine-readable text documents. Once digitized, the text from these documents can be easily searched, edited, stored, and managed, making it much more accessible and useful.
- Vehicle license plate identification.<sup>16</sup> OCR is a key technology behind Automatic Number Plate Recognition (ANPR) systems. These systems use OCR to read the license plate numbers of vehicles from digital images or video feeds for purposes like traffic enforcement, toll collection, or parking management.
- Consumer behavior and market analysis: extracting data from retail receipts and consumergenerated content (such as reviews or handwritten notes), which can then be analyzed for insights and consumer behavior analysis.<sup>17</sup>
- Transforming documents into text that can be read aloud to visually impaired or blind users.<sup>18</sup>
   OCR is used in assistive technologies to convert printed text into digital text, which can then be read aloud using text-to-speech (TTS) systems.
- Logistics and warehouse automation.<sup>19</sup> OCR can be used to automate processes such as inventory management, shipping, and receiving of goods. For example, OCR can be used to read labels, barcodes, or other identifiers on packages, enabling automatic tracking and sorting of goods.
- Medical Documentation Transcription & Automation.<sup>20</sup> OCR can be used to integrate paper and images originating from existing patient records into new electronic health records (EHR). It extracts the data required to automatically associate the information in a patient record, such as a medical record number, date of birth, patient name, etc., with the right electronic health record. It can also be used to digitize medical prescriptions.

A real use case of OCR in the medical sector is the ABBYY<sup>21</sup> medical records management software.

<sup>13</sup> Example: https://rossum.ai/

<sup>&</sup>lt;sup>14</sup> Example: https://www.klippa.com/en/blog/information/4-ways-to-perform-document-based-kyc-checks-with-ocr-and-ai/

<sup>&</sup>lt;sup>15</sup> Example: https://pdf.abbyy.com/finereader-pdf-for-mac/how-to/digitize-paper/

<sup>&</sup>lt;sup>16</sup> Example: https://platerecognizer.com/

<sup>&</sup>lt;sup>17</sup> Example: https://microblink.com/commerce/receipt-ocr/

<sup>&</sup>lt;sup>18</sup> Example: https://www.afb.org/blindness-and-low-vision/using-technology/assistive-technology-

videos/scanning-and-specialized-ocr

<sup>&</sup>lt;sup>19</sup> Example: https://www.innovapptive.com/blog/using-optical-character-recognition-ocr-to-overcome-

<sup>3-</sup>supply-chain-bottlenecks

<sup>&</sup>lt;sup>20</sup> Example: https://nanonets.com/blog/ocr-for-healthcare/

<sup>&</sup>lt;sup>21</sup> https://www.abbyy.com/solutions/healthcare/capture-to-ehr/

## 2. Data protection and privacy risk identification Definition of the criteria to consider when identifying risks and their categorization

To help identify risks associated to the use of data extraction technologies like OCR we can make use of a variety of risk factors.

Risk factors are conditions associated with a higher probability of undesirable outcomes. They can help to identify, assess, and prioritize potential risks. For instance, using health data and processing large volumes of data are risk factors with a high level of risk. Acknowledging them in your own use case, can help you identify related potential risks and their severity. In this case, an example of associated risk with a high severity could be 'a risk of violation of patients privacy due to a data breach'.

The risk factors shown below are the result of analysing the contents of legal instruments such as the GDPR<sup>22</sup>, the EUDPR<sup>23</sup>, the EU Charter<sup>24</sup> and other applicable guidelines related to privacy and data protection<sup>25</sup>.

The following risk factors can help us identify data protection and privacy <u>high level risks</u> in data extraction technologies like OCR:

| High level Risk / Important concerns                                     | Examples of applicability                                    |
|--|--|
| Sensitive & impactful purpose of the processing                          | - When documents are classified and archived                 |
| Using a OCR system to decide on or prevent the exercise of               | automatically, and the classification could have an impact   |
| fundamental rights of individuals, or about their access to a service,   | on the data subject. This could also apply in use cases such |
| the execution or performance of a contract, or access to financial       | as banking, health sector or ANPR <sup>26</sup> .            |
| services is a concern, especially if these decisions will be automated   |  |
| without human intervention. Wrong decisions could have an                |  |
| adverse impact on individuals.   |  |
| Processing sensitive data  | - When using OCB for digitizing medical records or legal     |
| When the OCR system is processing sensitive data such as: health         | documents by courts  |
| data special categories of data personal data related to convictions     | - When using OCR for digitizing invoices or other processes  |
| and criminal offences, financial data, behavioural data, unique          | in banking sector, for consumer behavior and market          |
| identifiers, location data, etc. This is a reason of concern since       | analysis, in data extraction from Identification documents   |
| processing inappropriately this personal data could negatively           | and bank cards. ANPR and the digitization of medical         |
| impact individuals.  | records.   |
| Large scale processing   | This could apply to most of OCR use cases since data         |
| Processing high volumes of personal data is a reason of concern,         | extraction technologies like OCR are usually applied to      |
| especially if these personal data are sensitive. The higher the          | large volumes of data.                                       |
| volume the bigger the impact in case of a data breach or any other       |  |
| situation that put the individuals at risk.                              |  |
| Processing data of vulnerable individuals                                | This could be the case when OCR solutions are used in the    |
| This is a concern because vulnerable individuals often require           | health sector, at schools, social services organizations,    |
| special protection. Processing their personal data without proper        | government institutions, employers, etc.                     |
| safeguards can lead to violations of their fundamental rights. Some      |  |
| examples of vulnerable individuals are children, elderly people,         |  |
| people with mental illness, disabled, patients, people at risk of social |  |
| exclusion, asylum seekers, persons who access social services,           |  |
| employees, etc.  |  |
|  |  |

<sup>&</sup>lt;sup>22</sup> General Data Protection Regulation (2016/679)

<sup>&</sup>lt;sup>23</sup> European Union Data Protection Regulation (Reg. 2018/1725)

<sup>&</sup>lt;sup>24</sup> Charter of Fundamental Rights of the European Union (2012/C 326/02)

<sup>&</sup>lt;sup>25</sup> Pag. 79, AEPD, "Risk Management and Impact Assessment in Processing of Personal Data", 2021

<sup>&</sup>lt;sup>26</sup> Automatic Number Plate Recognition

| Low data quality<br>The low data quality of the input data and/or the training data is a<br>concern bringing possible risks of inaccuracies in the generated   | OCR systems are not 100% accurate and quality issues in the input data are common.   |
|--|--|
| output what could cause wrong identification of characters and have other adverse impacts depending on the use case.   |  |
| Insufficient security measures<br>The lack of sufficient safeguards could be the cause of a data breach.<br>Data could also be transferred to states or organisations in other<br>countries without an adequate level of protection. | <ul> <li>This could be the case if there are not sufficient<br/>safeguards implemented to protect the input data and the<br/>results of the processing. This could be applicable to any<br/>use case.</li> <li>Data extraction technologies like OCR are often offered<br/>as SaaS solutions. Input data could be sent for processing<br/>to countries without an adequate level of protection.</li> </ul> |

## Presentation of examples of risks specific to OCR

Technologies for data extraction like OCR can present different types of privacy and data protection risks. The number and type of risks will depend on the use case, the context in which the technology is being applied as well as the different risks factors previously identified.

We are going to analyze different risks related to the procurement, development and use of this technology.

#### Data protection and privacy risks posed by the **procurement** of those types of AI systems:

Data extraction solutions are frequently available as SaaS solution from third party providers. Due to the different type of configurations available and the required maintenance of the models used, the use of an external supplier is usually the preferred option for users of this technology.

Some third party data extraction systems, though rarely, can also be hosted on-premises.

| Data Protection   | Risk description  | GDPR potential   | Examples   | Risk applicable                                     |
|---|---|--|--|---|
| and Frivacy Misks   |   | inpact   |  | provision   |
| Insufficient<br>protection of<br>personal data<br>that eventually<br>can be the cause<br>of a data breach                     | Safeguards for the<br>protection of personal<br>data are not<br>implemented or are<br>insufficient  | Infringement of Art.<br>32 Security of<br>processing, Art. 5 (f)<br>Integrity and<br>confidentiality and<br>Art. 9 Processing of<br>special categories of<br>personal data                           | OCR systems that process text<br>containing personal data could be<br>not properly secured. This could be<br>the case if for instance,<br>transmission of data is not secure,<br>data are not stored encrypted or<br>with an adequate access control<br>mechanism. | <ul> <li>SaaS cloud</li> <li>On-premises</li> </ul> |
| Possible adverse<br>impact on data<br>subjects that<br>could negatively<br>impact<br>fundamental<br>rights                    | The output of the<br>system could have an<br>adverse impact on the<br>individual if erroneous<br>data are used for<br>important decisions | Infringement of Art. 5<br>(d) Accuracy, Art. 5(a)<br>Fairness, Art.<br>22Automated<br>individual decision-<br>making, including<br>profiling, Art. 25 Data<br>protection by design<br>and by default | A system providing output that is<br>not accurate and does not provide<br>with mechanisms to amend errors.<br>Or when vendors claim their<br>system offers certain performance,<br>but this is not reproduced in real<br>cases.                                    | <ul> <li>SaaS cloud</li> <li>On-premises</li> </ul> |
| Lack of<br>compliance with<br>GDPR by not<br>granting data<br>subjects their<br>right to data<br>rectification and<br>erasure | Data subjects' requests<br>to rectify or to erase<br>personal data cannot<br>be completed   | Infringement of Art.<br>16 and Art. 17: Right<br>to rectification and<br>right to erasure  | A low-quality output could prevent<br>a controller from finding all the<br>data of a data subject in their data<br>storage since the data cannot be<br>matched properly.<br>This could also be the case If there<br>is not a possibility to search for the         | <ul> <li>SaaS cloud</li> <li>On-premises</li> </ul> |

|   |   |   | data subject's data in the output<br>and to correct and delete data.  |   |                           |
|---|---|---|---|---|---------------------------|
| Unlawful<br>repurpose of<br>personal data         | Personal data extracted<br>is used for a different<br>purpose                               | Infringement of Art. 5<br>(b) Purpose<br>limitation, Art. 5(a)<br>Lawfulness, fairness<br>and transparency,<br>Art. 29 Processing<br>under the authority<br>of the controller or<br>processor | This could be the case if the<br>supplier uses the input and/or<br>output data for training the ML<br>models without this being formally<br>agreed on beforehand.                                       | • | SaaS cloud                |
| Unlawful<br>unlimited storage<br>of personal data | Input data and/or data<br>extracted from images<br>is being stored longer<br>than necessary | Infringement of Art. 5<br>(e) Storage limitation  | The system could be unnecessarily<br>storing input data that is not<br>directly relevant to the OCR<br>process.<br>In some cases, the output could be<br>stored by the vendor longer than<br>necessary. | • | SaaS cloud<br>On-premises |
| Unlawful transfer<br>of personal data             | Data are being<br>processed in countries<br>without an adequate<br>level of protection      | Infringement of Art.<br>44 General principle<br>for transfers, Art. 45<br>Transfers on the basis<br>of an adequacy<br>decision, Art. 46<br>Transfers subject to<br>appropriate<br>safeguards  | Data extraction solutions could<br>store and be processing the data in<br>countries that do not offer enough<br>safeguards.   | • | SaaS cloud                |

Data protection and privacy risks posed by the **development** of those types of AI systems: The development of data extraction technologies can also face data protection and privacy risks. Risks could arise at different phases of the development life cycle, that is why it is important to implement an iterative process for the identification of this type of risks.

The development of an OCR system typically involves training machine learning models on large datasets of annotated images or documents. These datasets can consist of various types of digital and printed documents. The data used for training an OCR system typically includes:

- Training Data: this data includes a diverse set of images or documents that represent the target domain. It encompasses a wide range of fonts, text sizes, styles, layouts, and document types.
- Validation Data: a separate portion of the dataset is reserved for validation purposes during the model development process.
- Test Data: the other portion of the dataset that is used to evaluate the final performance of the trained OCR system. The test data should be representative of the real-world scenarios and provide a fair assessment of the system's accuracy and reliability.

OCR system developers often curate or collect their own datasets, which can include publicly available data, proprietary data, or datasets obtained through partnerships or collaborations. It is important to mention that training data can introduce certain risks in the development of OCR systems. Here are a few key considerations:

- Bias present in the training data, such as imbalances in document types, languages, or fonts, can impact the OCR system's performance and introduce unfairness.
- Inaccurate or incomplete annotations in the training data can adversely affect the performance of the OCR system. If the labeled data contains errors or inconsistencies, the

model may learn incorrect patterns or struggle to generalize well to unseen data. Ensuring high-quality annotations is crucial for effective training.

- If the training data does not adequately cover the full range of document types, fonts, text sizes, or languages encountered in real-world scenarios, the system may struggle to accurately recognize text in unseen or challenging conditions.
- The training data may contain sensitive or private information, such as personal details or confidential documents.
- Training data could be collected and used in an unethical manner, without respecting privacy, consent, copyright and other legal obligations.

The following table offers an overview of data protection and privacy risks that developers of OCR systems should consider during the design and development phase. The idea behind this table is to make developers conscious of privacy by design choices that can help prevent risks:

| Data Protection and   | Risk description  | GDPR Potential Impact  | Examples   |
|---|---|--|--|
| Privacy Risks   |   |  |  |
| Insufficient<br>protection of<br>personal data that<br>eventually can be the<br>cause of a data<br>breach               | Safeguards for the protection<br>of personal data that is part<br>of the training dataset are<br>not implemented or are<br>insufficient | Infringement of Art. 32<br>Security of processing, Art.<br>5 (f) Integrity and<br>confidentiality and Art. 9<br>Processing of special<br>categories of personal data                           | We could be using third party libraries,<br>SDK <sup>27</sup> or applications for the development<br>of the OCR system, and we could be<br>leaking data to these third parties. The<br>system could be integrated with other<br>systems internally and the transmission of<br>input data could be insecure; data could be<br>stored unencrypted and with inadequate<br>access control mechanisms. If using the<br>cloud, this could be not configured<br>according to security best practices. |
| Possible adverse<br>impact on data<br>subjects that could<br>negatively impact<br>fundamental rights                    | The output of the system<br>could have an adverse<br>impact on the individual if<br>inaccurate data are used for<br>important decisions | Infringement of Art. 5 (d)<br>Accuracy, Art. 5(a) Fairness,<br>Art. 22 Automated<br>individual decision-making,<br>including profiling, Art. 25<br>Data protection by design<br>and by default | Confidence levels could be based on<br>validation rules and historical data what<br>could act as a proxy for the quality of the<br>extraction. This could prevent errors in<br>documents for being flagged. OCR systems<br>might then assign high confidence scores<br>to incorrect predictions (false positives) or<br>low confidence scores to correct<br>predictions (false negatives) what can lead<br>to incorrect interpretations or<br>misclassifications.                              |
| Lack of compliance<br>with GDPR by not<br>granting data subjects<br>their right to data<br>rectification and<br>erasure | Data subjects' requests to<br>rectify or to erase personal<br>data cannot be completed  | Infringement of Art. 16 and<br>Art. 17 Right to rectification<br>and right to erasure  | A low-quality output could prevent a user<br>from finding all the data of a data subject<br>in their data storage since the data cannot<br>be matched properly. This is more<br>problematic if the developed application<br>does not provide with search output<br>format mechanisms and high accuracy<br>levels.  |
| Unlawful unlimited<br>storage of personal<br>data   | Input data and/or data<br>extracted from the images<br>are being stored longer than<br>necessary  | Infringement of Art. 5 (e)<br>Storage limitation   | This could be the case if training datasets<br>containing personal data are stored for too<br>long. But it could also be the case if the<br>system is developed in a way where input<br>and output data are automatically stored<br>without offering the user the possibility for<br>deletion.   |

<sup>&</sup>lt;sup>27</sup> SDK stands for software development kit. SDK is a set of software-building tools for a specific platform.

| Breach of the data | Extensive processing of        | Infringement of Art. 5 (c) | Certain OCR systems require large amounts              |
|--------------------|--------------------------------|----------------------------|--|
| minimization       | personal data for training the | Data minimisation          | of data to train the models <sup>28</sup> . Tasks that |
| principle          | model                          |                            | require handling more diverse fonts, styles,           |
|                    |                                |                            | or languages may generally require a larger            |
|                    |                                |                            | dataset to capture the necessary                       |
|                    |                                |                            | variability.   |

#### Data protection and privacy risks posed by the use of those types of AI systems:

Users of data extraction technologies need to consider the risks related to their specific use cases and context. Making use of the risk factors or evaluation criteria can facilitate the identification of those risks. For instance, the criteria 'large-scale processing of personal data' can already trigger the identification of risky processing activities that could result in harm.

When using an OCR solution, users have three different service model provisions available: SaaS solution from third party providers hosted in the cloud, third party solutions hosted on-premises<sup>29</sup> and self-developed own solutions hosted on- premises.

| Data Protection   | Risk description  | GDPR Potential   | Examples  | Risk applicable on   |  |
|---|---|--|---|--|--|
| and Privacy Risks   |   | Impact   |   | provision  |  |
| Insufficient<br>protection of<br>personal data that<br>eventually can be<br>the cause of a data<br>breach | Safeguards for the<br>protection of<br>personal data are<br>not implement or<br>are insufficient  | Infringement of Art.<br>32 Security of<br>processing, Art. 5 (f)<br>Integrity and<br>confidentiality, and<br>Art. 9 Processing of<br>special categories of<br>personal data                          | OCR systems that process text<br>containing personal data could be<br>not properly secured. This could be<br>the case if transmission of input<br>data is not secure, data are not<br>stored encrypted or with an<br>adequate access control<br>mechanism. This is especially<br>sensitive when we are processing<br>special category of personal data<br>when using OCR for digitizing<br>medical records, criminal data and<br>banking information. | <ul> <li>SaaS cloud</li> <li>Third party on-<br/>premises</li> <li>Self-developed</li> </ul> |  |
| Possible adverse<br>impact on data<br>subjects that could<br>negatively impact<br>fundamental rights      | The output of the<br>system could have<br>an adverse impact<br>on the individual if<br>erroneous data are<br>used for important<br>decisions. | Infringement of Art. 5<br>(d) Accuracy, Art. 5(a)<br>Fairness, Art. 22<br>Automated individual<br>decision-making,<br>including profiling,<br>Art. 25 Data<br>protection by design<br>and by default | Errors in the output could attribute<br>incorrectly actions to an individual<br>or group (misspelling errors in<br>names and dates for instance). This<br>could have especially a big impact<br>when using OCR for digitizing<br>medical records, banking ID<br>validation, legal documents with<br>sensitive information and criminal<br>records.  | <ul> <li>SaaS cloud</li> <li>Third party on-<br/>premises</li> <li>Self-developed</li> </ul> |  |
| Possible adverse<br>impact on data<br>subjects and lack of<br>compliance with<br>GDPR requirement         | Data subjects are<br>subjected to an<br>automatic decision-<br>making process<br>without human  | Infringement of Art.<br>22 Automated<br>individual decision-<br>making, including<br>profiling, Art. 9   | The output of an OCR system could<br>be used to make automatic<br>decisions which produce legal<br>effects or similarly significant<br>effects on data subjects, this could   | <ul> <li>SaaS cloud</li> <li>Third party on-<br/>premises</li> <li>Self-developed</li> </ul> |  |

<sup>&</sup>lt;sup>28</sup> The size of a training dataset can vary depending on multiple factors such as the complexity of the documents, the diversity of fonts and text styles, and the desired level of accuracy. For simpler document types with limited variations in fonts and layouts, a smaller training dataset may be sufficient to achieve reasonable accuracy. However, for more complex document types or scenarios requiring high accuracy, a larger and more diverse training dataset is typically necessary.
<sup>29</sup> Example: https://learn.microsoft.com/en-us/azure/applied-ai-services/form-recognizer/faq?view=form-recog-3.0.0

## AI Possible Risks & Mitigations - Optical Character Recognition (OCR)

| of providing human<br>intervention for<br>processing that can<br>have a legal or<br>important effect on<br>the data subject | intervention,<br>and/or there is a<br>processing of<br>special categories<br>of personal data | Processing of special<br>categories of personal<br>data   | be the case when using OCR in the<br>banking sector for identity<br>verification by enabling automatic<br>comparison and validation of the<br>provided data against reference<br>databases or identity records. Or<br>when data are extracted from legal<br>contracts or financial statements<br>and is then analyzed automatically<br>to identify non-compliant clauses,<br>suspicious activities, or anomalies,<br>triggering appropriate actions or<br>alerts.   |  |
|---|---|---|---|--|
| Lack of compliance<br>with GDPR by not<br>granting data<br>subjects their right<br>to data rectification<br>and erasure     | Data subjects'<br>requests to rectify<br>or to erase personal<br>data cannot be<br>completed  | Infringement of Art.<br>16 and Art. 17 Right<br>to rectification and<br>right to erasure  | This could be the case If there is<br>not a possibility to search for the<br>data subject's data in the output<br>and to correct and delete it.   | <ul> <li>SaaS cloud</li> <li>Third party on-<br/>premises</li> <li>Self-developed</li> </ul> |
| Unlawful unlimited<br>storage of personal<br>data   | Input data and/or<br>output data are<br>being stored longer<br>than necessary                 | Infringement of Art. 5<br>(e) Storage limitation  | In principle an OCR system doesn't<br>need to store the input data unless<br>it is necessary for audit,<br>verification, or archival purposes.<br>The system should avoid<br>unnecessary retention or storage of<br>input data that is not directly<br>relevant to the OCR process.<br>The storage of an OCR system's<br>output depends on the specific<br>application and requirements. In<br>some cases, the output could be<br>stored by the vendor longer than<br>necessary.<br>But this could also be the case on-<br>premises if we are not applying<br>data retention rules to our stored<br>data. | <ul> <li>SaaS cloud</li> <li>Third party on-premises</li> <li>Self-developed</li> </ul>      |
| Breach of the data<br>minimization<br>principle   | Extensive<br>processing of<br>personal data for<br>training the model                         | Infringement of Art. 5<br>(c) Data minimisation   | There are certain OCR systems that<br>might require the user to input<br>samples of data until the system<br>has learned the additional features.<br>The input data is used to train or<br>fine-tune the OCR model, enabling<br>it to handle the specific<br>complexities of the given OCR task.<br>This can be the case in OCR<br>systems used in medical fields that<br>may require specialized training on<br>medical terminology and<br>handwritten prescriptions, also for<br>handwritten recognition, when<br>using low-quality documents or<br>specialised fonts.                                  | <ul> <li>SaaS cloud</li> <li>Third party on-<br/>premises</li> <li>Self-developed</li> </ul> |
| Unlawful transfer of personal data  | Data are being<br>processed in<br>countries without<br>an adequate level<br>of protection     | Infringement of Art.<br>44, General principle<br>for transfers, Art. 45<br>Transfers on the basis<br>of an adequacy<br>decision, Art.46<br>Transfers subject to | OCR systems could store the data<br>and be processing the data in<br>countries that do not offer enough<br>safeguards.<br>This could also be the case with<br>self-developed systems if we store<br>the data in the cloud.  | <ul> <li>SaaS cloud</li> <li>Self-developed if<br/>using Cloud</li> </ul>                    |

|  | appropriate |  |
|--|-------------|--|
|  | safeguards  |  |

## 3. Data protection and privacy risk assessment

Once risks have been identified, it is time to proceed with their classification. The actual risk level or risk classification will depend on the specific use case and context.

The GDPR outlines in Recital 90 the importance of establishing the context: "taking into account the nature, scope, context and purposes of the processing and the sources of the risk".

This is an important process when performing a privacy risk assessment to manage risks to the rights and freedoms of natural persons.

The following processes are<sup>30</sup>:

- assessing the likelihood and severity of the risks;

- treating the risks by mitigating the identified risks and in that way ensuring the protection of personal data and demonstrating compliance with the GDPR and EUDPR.

There are different risk management methodologies available to classify and assess risks. It is not the purpose of this document to define or establish a methodology to be used since this is a choice that should be left to each organization. But for the purpose of this document, we will use the international standards that have been previously referenced in the WP29<sup>31</sup> and the AEPD<sup>32</sup> Guidelines.

In general risk management terms, risk can be summarized in one equation:

#### Risk = Likelihood x Severity

This means that risk is the probability of an event occurring, multiplied by the potential impact or severity incurred by the event.

To assess the level of risk of the data protection and privacy risks identified when procuring, developing and using data extraction technologies, we first need to estimate the likelihood and severity of the identified risks happening.

## Criteria to establish the likelihood of OCR risks. How to assess likelihood.

To determine the likelihood of the risks of data extraction technologies we are using the following four level risk classification matrix:

| Level of    | Likelihood Definition                 |
|-------------|---------------------------------------|
| Consequence |                                       |
| Very High   | High likelihood of an event occurring |

<sup>&</sup>lt;sup>30</sup> Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, Article 29 Data Protection Working Party, Last revision 2017

<sup>&</sup>lt;sup>31</sup> ISO 31000:2009, Risk management — Principles and guidelines, International Organization for Standardization (ISO); ISO/IEC 29134, Information technology – Security techniques – Privacy impact assessment – Guidelines, International Organization for Standardization (ISO).

<sup>&</sup>lt;sup>32</sup> ISO 31010:2019, Risk management — Risk Assessment Techniques, International Organization for Standardization (ISO)

High Substantial probability of an event occurring

Low probability of an event occurring

Unlikely There is no evidence of such a risk materializing in any case

Likelihood can only be determined based on specific risks and use cases. We will look later at a specific example to better understand how this process works.

## Criteria to establish the severity of OCR risks. How to assess severity.

To determine the severity of risks of data extraction technologies we are using the four level risk classification matrix<sup>33</sup>:

| Level of<br>Severity | Severity Definition   |
|----------------------|---|
| Very                 | It affects the exercise of fundamental rights and public freedoms, and its consequences are   |
| Significant          | irreversible and/or the consequences are related to special categories of data or to criminal |
|                      | offences and are irreversible and/or it causes significant social harm, such as               |
|                      | discrimination, and is irreversible and/or it affects particularly vulnerable data subjects,  |
|                      | especially children, in an irreversible way and/or causes significant and irreversible moral  |
|                      | or material losses.   |
| Significant          | The above cases when the effects are reversible and/or there is loss of control of the data   |
|                      | subject over their personal data, where the extent of the data are high in relation to the    |
|                      | categories of data or the number of subjects and/or identity theft of data subjects occurs    |
|                      | or may occur and/or significant financial losses to data subjects may occur and/or loss of    |
|                      | confidentiality of data subject or breach of the duty of confidentiality and/or there is a    |
|                      | social detriment to data subjects or certain groups of data subjects                          |
| Limited              | Very limited loss of control of some personal data and to specific data subjects, other than  |
|                      | special category or irreversible criminal offences or convictions                             |
|                      | and/or negligible and irreversible financial losses and/or loss of confidentiality of data    |
|                      | subject to professional secrecy but not special categories or infringement penalties          |
| Very Limited         | In the above case (limited), when all effects are reversible                                  |

The severity criteria are related to a loss of privacy that is experienced by the data subject but that may have further related consequences impacting other individuals and/or society.

## Example of OCR specific risk assessment

#### Use case: OCR system for the digitization of legal documents

**Scenario:** We want to digitize legal documents containing court filings for archiving purposes. The documents contain sensitive personal data such us criminal history, health and financial information. We do not have the expertise to develop and host ourselves an OCR system, so we are going to contract a third-party provider offering a SaaS solution in the cloud.

<sup>&</sup>lt;sup>33</sup> Pag. 77, AEPD, "Risk Management and Impact Assessment in Processing of Personal Data", 2021

The following risk factors/ important concerns from section 2 could be applicable in our specific use case:

| Risk factor                    | Use case applicability  |
|--------------------------------|---|
| Processing sensitive data      | Personal data related to convictions and criminal offences,   |
|                                | health and financial information                              |
| Large scale processing         | The volume of data to be processed is high                    |
| Processing data of vulnerable  | Criminal records  |
| individuals                    |   |
| Low data quality               | We do not know if the dataset is of sufficient quality        |
| Insufficient security measures | We might transfer personal data to states or organisations in |
|                                | other countries without an adequate level of protection.      |
|                                | There could be a possibility of a data breach. Third party    |
|                                | vendor solution has not been chosen yet; this must be taken   |
|                                | into account when making a choice.                            |
| Infringement of regulatory     | We might store the data too long, vendor might use it for     |
| requirements                   | other purposes than OCR, and we might not be able to answer   |
|                                | data subject rights requests for instance.                    |

Based on the identified risks factors we are going to identify together with other stakeholders<sup>34</sup> the data protection and privacy risks that could arise with the OCR implementation.

We are going to use as foundation the risks identified in section 2 for procurement, and we are going to assess what is the likelihood of the identified risks and assign to each risk one of the 4 likelihood classification levels from the matrix: Very high, High, Low, Unlikely.

| Data Protection  | Risk factor   | Risk description   | Likelihood | Reasoning  |
|--|---|--|------------|--|
| and Privacy Risks  |   |  |            |  |
| Insufficient<br>protection of<br>personal data<br>what eventually<br>can be the cause<br>of a data breach  | <ul> <li>Insufficient</li> <li>security</li> <li>measures</li> <li>Processing</li> <li>sensitive data</li> <li>Large scale</li> <li>processing</li> <li>Processing</li> <li>data of</li> <li>vulnerable</li> <li>individuals</li> </ul> | Att       Safeguards for the protection of personal data are not implemented or are insufficient       Low       The third-party s reviewed, have i measures such a strong access colencryption at res         Bata       Bata       Bata       Security processed |            | The third-party suppliers we have<br>reviewed, have implemented security<br>measures such as secure transmissions,<br>strong access control measures, and data<br>encryption at rest.<br>We, as user/customer have also strong<br>security processes implemented internally.                           |
| Possible adverse<br>impact on data<br>subjects that<br>could negatively<br>impact<br>fundamental<br>rights | <ul> <li>Low data</li> <li>quality</li> <li>Large scale</li> <li>processing</li> <li>Processing</li> <li>data of</li> </ul>   | The output of the system could<br>have an adverse impact on the<br>individual if erroneous data are<br>used for important decisions  | Low        | Though the impact would be high, we<br>could consider the likelihood of the risk<br>happening low due to the nature of the<br>processing for just the purpose of<br>archiving with not further analysis of the<br>data. This specific assessment could also<br>be done from the perspective of what is |

<sup>&</sup>lt;sup>34</sup> Meaningful involvement of different stakeholders during the risk assessment process: https://ecnl.org/sites/default/files/2023-

<sup>03/</sup>Final%20Version%20FME%20with%20Copyright%20%282%29.pdf

#### AI Possible Risks & Mitigations - Optical Character Recognition (OCR)

|   |   |  | 1   |  |
|---|---|--|---|--|
|   | vulnerable<br>individuals   |  |   | the likelihood of the output having a<br>negative impact just in general and in that<br>case, we could consider a higher<br>probability. It is important to assess the<br>risks based on context.  |
| Lack of<br>compliance with<br>GDPR by not<br>granting data<br>subjects their<br>right to data<br>rectification and<br>erasure | <ul> <li>Processing<br/>sensitive data</li> <li>Processing<br/>data of<br/>vulnerable<br/>individuals</li> <li>Infringement<br/>of regulatory<br/>requirements</li> </ul>   | Data subjects' requests to<br>rectify or to erase personal data<br>cannot be completed       | Low   | For this exercise we assume that in the<br>OCR SaaS solution we are going to use in<br>our use case, the data can be easily<br>searched once it is digitized. There is<br>always a small possibility that not all data<br>has been properly extracted and cannot be<br>found during the search function. The<br>system offers the possibility to delete data<br>from the output since output is available in<br>formats that are modifiable. |
| Unlawful<br>repurpose of<br>personal data   | <ul> <li>Processing</li> <li>sensitive data</li> <li>Processing</li> <li>data of</li> <li>vulnerable</li> <li>individuals</li> <li>Insufficient</li> <li>security</li> <li>measures</li> <li>Infringement</li> <li>of regulatory</li> <li>requirements</li> </ul> | Personal data extracted is used<br>for a different purpose                                   | Low<br>*This risk<br>will be<br>unlikely in<br>an on-<br>premises<br>solution | Except in a case of unlawful processing by<br>the third party provider, in principle is the<br>probability of the data extraction results<br>being used to retrain and fine-tune the<br>OCR model low. Most SaaS solutions<br>delete the data or offer you the possibility<br>to decide if you want to share the results<br>for that purpose.  |
| Unlawful<br>unlimited storage<br>of personal data   | <ul> <li>Processing<br/>sensitive data</li> <li>Processing<br/>data of<br/>vulnerable<br/>individuals</li> <li>Insufficient<br/>security<br/>measures</li> <li>Infringement<br/>of regulatory<br/>requirements</li> </ul>   | Input data and/or data<br>extracted from images are<br>being stored longer than<br>necessary | Low   | For our use case we could consider for<br>instance, vendors that offer a 24 hours<br>automated deletion period what already<br>reduces the likelihood of this risk.<br>Vendors usually offer a 24 or max. 48<br>hours automated deletion period in SaaS<br>solutions. Self-developed OCR systems can<br>be configured in a way that input and<br>output can be immediately deleted or at a<br>scheduled moment.                              |
| Unlawful transfer<br>of personal data   | <ul> <li>Insufficient<br/>security</li> <li>measures</li> <li>Processing<br/>sensitive data</li> <li>Processing<br/>data of</li> <li>vulnerable</li> <li>individuals</li> <li>Infringement</li> <li>of regulatory</li> <li>requirements</li> </ul>                | Data are being processed in<br>countries without an adequate<br>level of protection          | Low   | In our specific use case, we have decided<br>to work with vendors that offer an<br>adequate level of protection what already<br>reduces the likelihood though not the<br>impact.   |

After the likelihood assessment, we are going to assess what is the impact of the identified risks on the data subjects, individuals and society. Based on that impact/severity assessment, we will assign one of the 4 severity classification levels: Very significant, Significant, Limited, Very limited.

## AI Possible Risks & Mitigations - Optical Character Recognition (OCR)

| Data Protection and<br>Privacy Risks  | Risk description   | Likelihood   | Severity                           | Reasoning   |
|---|--|--|------------------------------------|---|
| Insufficient<br>protection of<br>personal data what<br>eventually can be<br>the cause of a data<br>breach   | Safeguards for the protection of<br>personal data are not<br>implemented or are insufficient   | Low  | Very<br>significant                | The documents contain very sensitive<br>information, and a data breach could<br>cause significant harm to the data<br>subjects.   |
| Possible adverse<br>impact on data<br>subjects that could<br>negatively impact<br>fundamental rights  | The output of the system could<br>have an adverse impact on the<br>individual if erroneous data are<br>used for important decisions                  | Low  | Very<br>significant                | Using the system for other purposes<br>beyond archiving could have an<br>adverse impact on data subjects if the<br>output is inaccurate. This could be the<br>case if we use the system for search,<br>analyze, and retrieve of information.  |
| Lack of compliance<br>with GDPR by not<br>granting data<br>subjects their right<br>to data rectification<br>and erasure<br>Unlawful repurpose<br>of personal data | Data subjects' requests to<br>rectify or to erase personal data<br>cannot be completed<br>Personal data extracted is used<br>for a different purpose | Low<br>Low   | Significant<br>Very<br>significant | Not being able to rectify incorrect or<br>not up to date information could have<br>a significant impact on the data subject<br>due to the nature of the data being<br>digitized: criminal history, medical and<br>financial information.<br>This could have a big impact on the<br>data subjects if for instance the vendor |
|   |  | will be<br>unlikely in<br>an on-<br>premises<br>solution |                                    | output data and uses this afterwards<br>for non-agreed purposes especially due<br>to the nature of the personal data<br>contained in the documents.   |
| Unlawful unlimited<br>storage of personal<br>data   | Input data and/or data<br>extracted from images are being<br>stored longer than necessary  | Low  | Significant                        | An unlimited or unlawful storage of<br>personal data would worsen any data<br>breach affecting stored data.<br>And although if data are properly<br>protected while being stored would<br>limit the harm cause to the data<br>subject, it will still be an infringement<br>of the GDPR.                                     |
| Unlawful transfer of personal data  | Data are being processed in<br>countries without an adequate<br>level of protection  | Low  | Significant                        | Transferring the data to a country that<br>doesn't offer enough safeguards could<br>bring significant risks to the data<br>subjects.  |

## 4. Data protection and privacy risk treatment

#### Risk treatment criteria

i.e., mitigate, transfer, avoid or accept a risk.

The assessments of likelihood and severity will offer us the basis to obtain the risk level classification. Based on the four level classification used for likelihood and severity, we can use a matrix like the following to obtain the resulting final risk level classification: Very High, High, Medium, Low.

|            | Very High | Medium       | High    | Very high   | Very high   |
|------------|-----------|--------------|---------|-------------|-------------|
| Likelihood | High      | Low          | High    | Very high   | Very high   |
|            | Low       | Low          | Medium  | High        | Very high   |
|            | Unlikely  | Low          | Low     | Medium      | Very high   |
|            |           | Very limited | Limited | Significant | Very        |
|            |           |              |         |             | Significant |
|            |           |              | Seve    | eritv       |             |

Based on this matrix we can classify the risks identified in our use case as follows:

| Data Protection and Privacy<br>Risks  | Risk description  | Likelihood  | Severity            | Risk Level |
|---|---|---|---------------------|------------|
| Insufficient protection of<br>personal data what eventually<br>can be the cause of a data<br>breach               | Safeguards for the protection of personal data are not implemented or are insufficient  | Low   | Very<br>significant | Very High  |
| Possible adverse impact on<br>data subjects that could<br>negatively impact<br>fundamental rights                 | The output of the system could have an adverse<br>impact on the individual if erroneous data are used<br>for important decisions. | Low   | Very<br>significant | Very High  |
| Lack of compliance with GDPR<br>by not granting data subjects<br>their right to data rectification<br>and erasure | Data subjects' requests to rectify or to erase personal data cannot be completed  | Low   | Significant         | High       |
| Unlawful repurpose of personal data   | Personal data extracted is used for a different purpose   | Low<br>*This risk<br>will be<br>unlikely in<br>an on-<br>premises<br>solution | Very<br>significant | Very High  |
| Unlawful unlimited storage of personal data   | Input data and/or data extracted from images are<br>being stored longer than necessary  | Low   | Significant         | High       |
| Unlawful transfer of personal data  | Data are being processed in countries without an adequate level of protection   | Low   | Significant         | High       |

We have identified three risks with a very high level, and three with a high level. Best practices in risk management suggest that the mitigation of very high and high level risks should be prioritized.<sup>35</sup> The next step involves the implementation of a risk treatment plan.

<sup>&</sup>lt;sup>35</sup> https://www.pmi.org/learning/library/high-risk-critical-path-projects-7675

Risk treatment involves developing options for mitigating the risks and preparing and implementing action plans. The appropriate treatment option should be chosen on a contextual basis and considering a feasibility analysis<sup>36</sup> like the following:

- Evaluate the type of risk and the available mitigation measures that can be implemented.
- Compare the potential benefits gained from implementing the mitigation against the costs and efforts involved.
- Assess the impact on the purpose that is being pursued by implementing the OCR system.
- Evaluate what could be the reasonable expectations of individuals.
- Assess the impact mitigation measures could have on transparency and fairness of the processing.

An analysis of these criteria is essential to risk mitigation and risk management planning and helps in determining whether the risk mitigation is justifiable.

The most common risk treatment criteria are: **Mitigate, Transfer, Avoid** and **Accept**. For each risk one of the criteria options will be selected:

- ✓ Mitigate Identify ways to reduce the likelihood or the severity of the risk.
- ✓ Transfer Make another party responsible for the risk (buy insurance, outsourcing, etc.).
- ✓ Avoid Eliminate the risk by eliminating the cause.
- ✓ Accept Nothing will be done.

Deciding whether a risk can be mitigated involves assessing the nature of the risk, understanding its potential impact, and evaluating potential mitigation measures such as implementing controls, adopting best practices, modifying processes, and using tools that can help reduce the likelihood or severity of the risk.

Not all risks can be fully mitigated. Some risks may be inherent and cannot be entirely avoided. In such cases, the goal is to reduce the risk to an acceptable level or to put in place measures that help manage the severity of the risk effectively.

#### Presentation of mitigation measure examples/risk treatment options

including an assessment of their practical feasibility and a definition of the criteria to define the level of mitigation obtained.

In our use case we have identified several very high and high level risks. After going through the feasibility analysis and the treatment criteria, we have decided that we cannot **transfer** the risks to any third party, we cannot **avoid** all the risks, and **acceptance** of the risks is an unacceptable option for us. As long as there are measures that we can implement to help us mitigate the risks, resulting in acceptable conditions to go on with the implementation, we choose the treatment option of risk **mitigation**.

We have identified the following risk mitigation measures:

<sup>&</sup>lt;sup>36</sup> "Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR", CIPL GDPR Interpretation and Implementation Project, 2016

| Data Protection and<br>Privacy Risks   | a Protection and Risk Level Risk Mitigation measures Feasibility Assessment<br>vacy Risks |   | New risk<br>Level after<br>mitigation  |                   |
|--|---|---|--|-------------------|
| Insufficient<br>protection of<br>personal data what<br>eventually can be the<br>cause of a data<br>breach:<br>safeguards for the<br>protection of<br>personal data are not<br>implemented or are<br>insufficient                                       | Very High   | <ul> <li>The third-party vendor chosen must have implemented security measures such as secure transmissions, strong access control measures, and data encryption at rest and sufficient privacy design strategies<sup>37</sup> to protect the data. We will ask certifications<sup>38</sup> and results of a pentest<sup>39</sup> to the vendor.</li> <li>As controller we can also protect the specific sensitive data in the documents by applying pseudonymization or anonymization techniques after the data extraction. Depending on the different needs, we could implement default anonymization or reversible data masking techniques, for instance allowing access to the unmasked data to certain people. If the OCR SaaS solution does not offer the possibility to implement these techniques<sup>40</sup>, we could decide to look for a vendor that can offer them or implement</li> </ul>  |  | Low               |
| Possible adverse<br>impact on data<br>subjects that could<br>negatively impact<br>fundamental rights:<br>the output of the<br>system could have an<br>adverse impact on<br>the individual if<br>erroneous data are<br>used for important<br>decisions. | Very High   | <ul> <li>1.In this specific use case, digitized documents are going to be archived and not been used for any other processing what would already mitigate this risk.</li> <li>2.But in cases where search and analysis of data is required, we would need to make sure that the OCR system we use offers a high percentage of accuracy guarantee. This is usually between 98-99%.<sup>41</sup> This accuracy should not only be at page level but also at character and word level what is often challenging. It is also important to follow best practices<sup>42</sup> when using OCR systems: making sure that our original documents have a good quality, considering things such as resolution, brightness, straightness, and discoloration before we scan text.</li> <li>Use of special fonts and low contrast could also affect accuracy. Another important aspect is that currently there are no systems offering 100% accuracy, and the only way to achieve that and avoid any error is by doing a human review and correction of the output.</li> </ul> | <ol> <li>Cost of<br/>implementation:<br/>They highest cost is the<br/>effort that implies doing<br/>the human review when<br/>needed and providing the<br/>human resources for that.</li> <li>Impact on purpose of<br/>digitization and<br/>archiving: No</li> <li>Impact on<br/>expectations of<br/>individuals: No for<br/>data subjects, but it<br/>has an impact on the<br/>employees that would<br/>be in charge of the<br/>human review</li> <li>Impact on<br/>transparency and<br/>fairness of the<br/>processing: No if it is<br/>properly implemented<br/>and information about<br/>how the accuracy of<br/>the system works is<br/>provided to users and<br/>eventually to data<br/>subjects.</li> </ol> | 1.Low<br>2.Medium |

<sup>37</sup> "Privacy Design Strategies" Jaap-Henk Hoepman, 2022

<sup>38</sup> Security certifications such as ISO27001, and SOC2

 <sup>&</sup>lt;sup>39</sup> A penetration test, colloquially known as a pentest or ethical hacking, is an authorized simulated cyberattack on a computer system, performed to evaluate the security of the system.
 <sup>40</sup> https://anonimiseren-bnas.nl/biqe-anonymization-2/
 <sup>41</sup> https://www.docsumo.com/blog/ocr-accuracy
 <sup>42</sup> https://www.docsumo.com/blog/ocr-accuracy

<sup>&</sup>lt;sup>42</sup> https://guides.library.illinois.edu/OCR/bestpractices

| Unlawful repurpose<br>of personal data:<br>personal data<br>extracted is used for<br>a different purpose   | Very High | <ul> <li>1.One of the best mitigation measure<br/>would be using a SaaS solution that offers<br/>the option to keep the data on premises.<br/>This is the case if the data processing takes<br/>place at location (where the OCR machine<br/>is located for instance) and the input is<br/>automatically deleted and the output data<br/>is only stored at the user location.</li> <li>2.If the data extraction takes place at the<br/>vendor's location, then a minimum of<br/>security measures such as access control,<br/>audit trail and encryption together with<br/>proper data protection agreements need<br/>to be in place.</li> </ul> | <ol> <li>Cost of<br/>implementation:</li> <li>If the on-premise option is<br/>offered by the vendor this<br/>could have an additional<br/>cost. It could also imply<br/>that we need to make<br/>resources available for<br/>taking care of the on-<br/>premise solution.</li> <li>Impact on purpose of<br/>digitization and<br/>archiving: No</li> <li>Impact on<br/>expectations of<br/>individuals: No</li> <li>Impact on<br/>transparency and<br/>fairness of the<br/>processing: No</li> </ol> | 1.Low<br>2.Medium |
|--|-----------|--|---|-------------------|
| Lack of compliance<br>with GDPR by not<br>granting data<br>subjects their right to<br>data rectification and<br>erasure:<br>Data subjects'<br>requests to rectify or<br>to erase personal<br>data cannot be<br>completed | High      | We could implement an OCR system that<br>offers editable output format. This will<br>allow as to search and edit text in the<br>output. This is not possible in other<br>formats like only searchable output<br>formats. This editable output format will<br>allow us to look up for the data subject's<br>information and update it or delete it.   | <ol> <li>Cost of<br/>implementation:</li> <li>If the option is offered by<br/>the vendor this could have<br/>an additional cost.</li> <li>Impact on purpose of<br/>digitization and<br/>archiving: No</li> <li>Impact on<br/>expectations of<br/>individuals: No</li> <li>Impact on<br/>transparency and<br/>fairness of the<br/>processing: Yes, on a<br/>positive way</li> </ol>  | Low               |
| Unlawful transfer of<br>personal data:<br>Data are being<br>processed in<br>countries without an<br>adequate level of<br>protection  | High      | We can implement an OCR system from a<br>third-party provider that is located in a<br>country offering adequate level of<br>protection.  | <ol> <li>Cost of<br/>implementation:</li> <li>This measure should in<br/>principle not have any<br/>additional cost, but it<br/>depends on the vendors<br/>availability.</li> <li>Impact on purpose of<br/>digitization and<br/>archiving: No</li> <li>Impact on<br/>expectations of<br/>individual: No</li> <li>Impact on<br/>transparency and<br/>fairness of the<br/>processing: No</li> </ol>   | Low               |
| Unlawful unlimited<br>storage of personal<br>data:<br>Input and output<br>data are being stored<br>longer than necessary   | High      | We could try to implement an OCR system<br>in which deletion of data can be configured<br>so that input and output data are deleted<br>from the system immediately after the<br>data extraction or at a scheduled moment<br>(this is by most vendor a period of 24 to 48<br>hours).<br>We could also implement a retention<br>period for the output data that we want to   | <ol> <li>Cost of<br/>implementation:</li> <li>If the option is offered by<br/>the vendor this could have<br/>an additional cost.</li> <li>Impact on purpose of<br/>digitization and<br/>archiving: No</li> </ol>  | Low               |

| archive and that is already in our own<br>premises.<br>If the vendor is storing the input and/or<br>output data, we should negotiate<br>contractual agreements about the | 3.<br>4. | Impact on<br>expectations of<br>individuals: No<br>Impact on<br>transparency and |
|--|----------|--|
| retention period of the data.  |          | fairness of the processing: No   |

#### Residual risk acceptance

After the feasibility assessment has been done and the mitigation measures have been identified and implemented, we should assess again the likelihood and severity of each risk to obtain a new risk classification level and in this way assess if there is any remaining or residual risk.

In our use case, after the assessment, all the risks have been reduced to the lowest risk level 'low', and there are two risks with a possible classification of 'Medium' depending in the example on the mitigation measure adopted.

We calculate the residual risk by evaluating the likelihood and severity of the risks that still exists despite the implemented mitigation measures. This residual risk represents the level of risk that remains after taking mitigation actions.

Once residual risk has been identified, we need to decide whether the residual risk is within acceptable levels for our organization. If it is, we can decide to accept it. If it's not, we would need to consider further mitigation strategies.

Some organizations establish criteria for the acceptability of residual risks based on elements such us social norms, benefits, harms, similar use cases, etc<sup>43</sup>.

Organizations must be able to justify their risk mitigation and acceptance decisions as part of their accountability obligations which also fall under the GDPR principle of accountability (Article 5.2, Recital 74).

## Example of general mitigation measures related to risks of OCR systems:

Choosing appropriate mitigation measures should be done on a case-by-case basis. We are going to examine some of the possible mitigation measures that could be implemented to mitigate privacy and data protection risks specific for data extraction technologies. These measures are general and not related to any specific use case.

| Data Protection and Privacy Risks          | Mitigation measures examples   |
|--|--|
| Insufficient protection of personal data   | As user, procurement entity and developer, it is important to verify <sup>44</sup> that APIs |
| what eventually can be the cause of a data | are securely implemented, transmission of data are protected with the adequate               |
| breach                                     | encryption protocols, data at rest is encrypted, there is an adequate access                 |
|  | control mechanism implemented, there are measures implemented for                            |
|  | protection and identification of insider threats, measures to mitigate supply chain          |
|  | attacks that could give access to the training data and/or the data storage and              |
|  | encryption keys, measures implemented to prevent risks associated to the use of              |
|  | deep learning such as the risk of reprogramming deep neural net attacks <sup>45</sup> ,      |

<sup>&</sup>lt;sup>43</sup> https://www.sciencedirect.com/topics/engineering/residual-risk

<sup>&</sup>lt;sup>44</sup> This could be done by performing a pentest and/or requesting pentest results to the vendor.

<sup>&</sup>lt;sup>45</sup> Elsayed et al , "Adversarial reprogramming of neural networks", 2018

|   |   | membership inference <sup>46</sup> , inversion <sup>47</sup> and poisoning attacks <sup>48</sup> . Also access and |  |  |  |
|---|---|--|--|--|--|
|   |   | change logs should be established to document access and changes to digitized                                      |  |  |  |
|   |   | records.   |  |  |  |
| ľ | Possible adverse impact on data subjects    | As <b>user</b> , implement OCR solutions that offer a minimum 98-99% accuracy. Often                               |  |  |  |
|   | that could negatively impact fundamental    | the systems offer the results of this metric after every data extraction. It is                                    |  |  |  |
|   | rights                                      | important to monitor the values and make the necessary adjustments and   |  |  |  |
|   |   | corrections to the results. Make sure the system recognizes different conditions                                   |  |  |  |
|   |   | annlicable to the input data   |  |  |  |
|   |   | The quality of input data is important. This is important for <b>users</b> of OCR system                           |  |  |  |
|   |   | as well as for <b>developers</b> that pood to use training data of quality to train their                          |  |  |  |
|   |   | as well as for developers that need to use training data of quality to train then                                  |  |  |  |
|   |   | Induces. As <b>developer</b> there are techniques and tools <sup>to</sup> you can use to reduce the                |  |  |  |
|   |   | Tow quality of input data such as binarization, deskweing <sup>55</sup> , rotation, increasing of                  |  |  |  |
|   |   | reducing brightness, scaling the image, or removing specific objects to improve                                    |  |  |  |
|   |   | the accuracy levels.   |  |  |  |
|   |   | Some of these techniques are also available in the configuration settings of OCR                                   |  |  |  |
| ŀ |   | third party solutions that are available as SaaS <sup>31</sup> solution and on-premises <sup>32</sup> .            |  |  |  |
|   | Possible adverse impact on data subjects    | <b>Users</b> could implement a human review process to verify the correctness of the                               |  |  |  |
|   | and lack of compliance with GDPR            | personal data, especially if this is sensitive, and a process to approve high risk                                 |  |  |  |
|   | requirement of providing human              | decisions after human verification has been done.  |  |  |  |
|   | intervention for processing that can have a | It is also important that the system provides with an overview of the accuracy and                                 |  |  |  |
|   | legal or important effect on the data       | the confidence levels achieved after the data extraction and with a dashboard or                                   |  |  |  |
|   | subject                                     | other type of interface for manual human review and correction.  |  |  |  |
|   |   | In certain use cases it might be necessary to implement a redress mechanism for                                    |  |  |  |
| L |   | data subjects.   |  |  |  |
|   | Lack of compliance with GDPR by not         | As user, developer and procurement entity, implement searchable and editable                                       |  |  |  |
|   | granting data subjects their right to data  | output format functions to identify the personal data in the data extracted so that                                |  |  |  |
|   | rectification and erasure                   | it is possible to respond to data erasure and rectification requests.  |  |  |  |
| I | Unlawful unlimited storage of personal      | As user and procurement entity make agreements with the third-party supplier                                       |  |  |  |
|   | data  | about how long the input data and output data should be stored. This can be part                                   |  |  |  |
|   |   | of the service contract, product documentation <sup>53</sup> or data processing agreement.                         |  |  |  |
|   |   | If data are being stored on your premises, establish retention rules and /or a                                     |  |  |  |
|   |   | mechanism for the deletion of data.  |  |  |  |
| I | Breach of the data minimization principle   | For <b>users</b> and <b>developers</b> , one possible way to mitigate this risk is by providing                    |  |  |  |
|   |   | documents to the OCR model where personal data has been replaced by synthetic                                      |  |  |  |
|   |   | data.  |  |  |  |
|   |   | As user, it is also important to compare the different OCR solutions available on                                  |  |  |  |
|   |   | the market to understand which systems require less volume of data to train the                                    |  |  |  |
|   |   | models and to improve the accuracy levels  |  |  |  |
| ŀ | Unlawful transfer of personal data          | As user and procurement entity, verify with the vendor where the data  |  |  |  |
|   |   | processing is taking place. Make the percessary safeguard diligences and when                                      |  |  |  |
|   |   | necessary perform a Data Transfer Impact Ascessment Make the necessary   |  |  |  |
|   |   | contractual agreements. Consider this risk when making a selection among   |  |  |  |
|   |   | different venders  |  |  |  |
| T |   |  |  |  |  |

Once risk mitigation measures have been implemented, it is crucial to continuously monitor their effectiveness. Implementing methodologies like threat modeling for the identification of risks,

<sup>&</sup>lt;sup>46</sup> Shokri et al, "Membership Inference Attacks Against Machine Learning Models", 2017

<sup>&</sup>lt;sup>47</sup> Zhang et al, "Generative Model-Inversion Attacks Against Deep Neural Networks", 2020

<sup>&</sup>lt;sup>48</sup> Junfeng Guo, Cong Liu, "Practical Poisoning Attacks on Neural Networks", 2020

<sup>&</sup>lt;sup>49</sup> See next section under 'Resources about how to improve OCR accuracy applying different tools and techniques'

<sup>&</sup>lt;sup>50</sup> Deskewing is the process of straightening an image that has been scanned or written crookedly. It is a process whereby skew is removed by rotating an image by the same amount as its skew but in the opposite direction.

<sup>&</sup>lt;sup>51</sup> https://help.abbyy.com/en-us/flexicapture/12/distributed\_administrator/template\_properties/

<sup>52</sup> https://selectec.com/on-premise-ocr/

<sup>&</sup>lt;sup>53</sup> In this document from Tencent Cloud in page 6, there is a retention Policy indicated for the images uploaded (input data) and the returned results (output): data is deleted upon completion of the processing https://main.qcloudimg.com/raw/document/intl/product/pdf/tencent-cloud\_1005\_50443\_en.pdf

maintaining a risk register and assigning risk owners are effective strategies for regularly reviewing and reassessing the risk landscape. This ensures that the implemented risk mitigation measures remain relevant and effective in preventing data protection and privacy risks that could adversely impact individuals and organizations.

## Reference to specific technologies, tools, methodologies, processes or strategies.

Unless standardised and freely and easily accessible, explanation on how these technologies, tools, methodologies and processes work.

#### Methodologies for measuring accuracy in OCR:

When evaluating the accuracy and quality of OCR results, there are various methods and metrics to consider. Character error rate (CER) and word error rate (WER) are quantitative metrics that measure the percentage of characters or words that are incorrectly recognized by the OCR system. Layout error rate (LER) is a qualitative metric that measures the degree of deviation between the OCR output and the original image in terms of layout, structure, and formatting.

To measure the extent of errors<sup>54</sup> between two text sequences we can use the Levenshtein distance metric to measure the difference between two string sequences. This is the minimum number of single-character (or word) edits (i.e., insertions, deletions, or substitutions) required to change one word (or sentence) into another.

#### Character Error Rate (CER)

CER calculation is based on the concept of Levenshtein distance, where we count the minimum number of character-level operations required to transform the reference text (aka ground truth) into the OCR output. It is represented with this formula:

CER=(S+D+I)/N

where:

S = Number of Substitutions

D = Number of Deletions

I = Number of Insertions

N = Number of characters in reference text

The output of this equation represents the percentage of characters in the reference text that was incorrectly predicted in the OCR output. The lower the CER value (with 0 being a perfect score), the better the performance of the OCR model. CER is relevant for extraction of particular sequences (e.g., social security number, phone number, etc.)

There is not a benchmark available for defining a good CER value, as it is highly dependent on the use case, the different scenarios and complexity. Some research studies<sup>55</sup> propose a good OCR accuracy should be CER 1-2% (i.e., 98-99% accurate)<sup>56</sup>.

Good OCR accuracy: CER 1-2% (i.e. 98–99% accurate) Average OCR accuracy: CER 2-10% Poor OCR accuracy: CER >10% (i.e. below 90% accurate)

#### Word Error Rate (WER)

Word Error Rate is relevant for the extraction of paragraphs and sentences of words with meaning (e.g., pages of books, newspapers). The formula for WER is the same as that of CER, but WER

<sup>&</sup>lt;sup>54</sup> "Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing", Nguyen, Adam, Coustaty, Nguyen, Doucet, 2019

<sup>&</sup>lt;sup>55</sup> <u>https://www.researchgate.net/publication/281583162</u> <u>Performance</u> <u>Comparison of OCR\_Tools</u> <sup>56</sup> <u>https://www.docsumo.com/blog/ocr-accuracy</u>

operates at the word level instead. It represents the number of word substitutions, deletions, or insertions needed to transform one sentence into another.

While CER and WER are handy, they are not bulletproof performance indicators of OCR models. This is because the quality and condition of the original documents (e.g., handwriting legibility, image DPI, etc.) plays a very important role and not just the OCR model itself.

Available tools for measuring CER and WER: CER metric: <u>https://huggingface.co/spaces/evaluate-metric/cer</u> WER metric: <u>https://huggingface.co/spaces/evaluate-metric/wer</u> WER: <u>https://www.amberscript.com/en/wer-tool/</u>

## OCR quality standard:

The ISO standard ISO/IEC 30116:2016 can be useful to evaluate the quality of the character recognition and data extraction output. The standardization project also defines test methods to evaluate OCR document quality.

 ISO/IEC 30116:2016 Information technology — Automatic identification and data capture techniques — Optical Character Recognition (OCR) quality testing <u>https://www.iso.org/obp/ui/#iso:std:iso-iec:30116:ed-1:v1:en</u>

#### Guidances:

- Digitalization Guide for Records with Long-term Retention at NYC Agencies <u>https://www.nyc.gov/assets/records/pdf/Digitization%20Guide%20for%20Records%20with%20</u> <u>Long%20Term%20Retention%20at%20NYC%20Agencies%2020161031.pdf</u>
- On-premise or Cloud OCR: A guide to help you decide: <u>https://www.klippa.com/en/blog/information/on-premise-ocr/</u>

## Resources about how to improve OCR accuracy applying different tools and techniques:

- Improve OCR accuracy using advanced pre-processing techniques
   <u>https://www.nitorinfotech.com/blog/improve-ocr-accuracy-using-advanced-preprocessing-techniques/</u>
- Tesseract: Improving the quality of the output <u>https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html</u>
- 15 tools for document Deskewing and Dewarping <u>https://safjan.com/tools-for-doc-deskewing-and-dewarping/</u>
- Improve the quality of your OCR information extraction <u>https://aicha-fatrah.medium.com/improve-the-quality-of-your-ocr-information-extraction-ebc93d905ac4</u>

#### Privacy preserving OCR techniques and tools:

 Confidential Optical Character Recognition Service with Cape <u>https://capeprivacy.com/blog/confidential-Optical-character-recognition-service-with-cape/</u>  Blur out Text in Images Using OCR in Next.js <u>https://cloudinary.com/blog/guest\_post/blur-out-text-in-images-using-ocr-in-next-js</u>

Methodologies and tools for the identification of data protection and privacy risks:

- Privacy Library of Threats (PLOT4ai) is a threat modeling methodology for the identification of risks in AI systems. It also contains a library with more than 80 risks specific to AI systems: <u>https://plot4.ai/</u>
- MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems), is a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems: <u>https://atlas.mitre.org/</u>
- Assessment List for Trustworthy Artificial Intelligence (ALTAI) is a checklist that guides developers and deployers of AI systems in implementing trustworthy AI principles: <u>https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificialintelligence-altai-self-assessment</u>

