Study Title	**The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas**
Working group code	Creation and Use: CU05
Document type	Final report
Status	Final version; Public
Version	12.0
Writers	Stefano Allegrezza, Mariella Guercio, Maria Mata Caravaca, Massimiliano Grandi, Bruna La Sorda
Date	November 1, 2023

Table of contents

1 SCOPE AND PURPOSE OF THE STUDY

1.1 Goals and introductory remarks

The overall goal of this study is to investigate the ability of AI to support creation (or recreation) of archival aggregations in order to address the issue of non-aggregated, unarranged, or de-contextualized records (both in the current and semi-current phases of their lifecycle). In many public administrations and private companies, documents are neither classified nor aggregated. In other cases, aggregations of documents are not properly created, which results in an uncontrolled number of documents that are unsorted, misplaced, and hard to find. In many cases metadata elements - necessary to ensure the reliability, trustworthiness, quality and sustainability of appraisal and acquisition - are missing. Despite progress on various technologies to support records management, current software products can only give limited help to carry out those activities. So, the research question this study aims to answer is: Can AI tools help to build or recreate archival aggregations and create metadata schemas for them?

For instance, consider the case of email management, which is one of the most time-consuming activities in both the public and in the private sector as well as in the organization of personal documents. People spend a lot of time to browse and read emails (and, obviously, answer them), to classify and file them, to appraise and perform many other repetitive tasks. So, we wondered if AI technologies could be useful for the effective automatic or semi-automatic management of emails, and in particular for classification and arrangement, filtering, association with metadata elements to describe the context of their creation and use, as well as setting-up automatic answering functions and automated appraisal and disposal.

Why is this topic so relevant? *Si parva licet componere magnis*, our answer could have the same motivation at the basis of the US update of the national research on artificial intelligence and the strategic plan development:

> *The federal government must place people and communities at the center by investing in responsible R&D that serves the public good, protects people's rights and safety, and advances democratic values. This update to the National AI R&D Strategic Plan is a roadmap for driving progress toward that goal.*[1]

Archivists have no doubts that the archival function serves (even if played in the private sector) the public good, protects people's rights and safety, and advances democratic values. Since the archival function is strongly based on the original documentary relations, their correct definition and maintenance, the recognition of relevance must be extended to the archival aggregations and to methods and tools in place for their qualified and efficient creation, and for ensuring their persistence.

This is more crucial in the digital environment, specifically when we often are helpless witnesses of the constant loss of the correct use of good practices in this domain and for this specific objective. The classification plans and filing systems based on functional classification are less and less applied as the massive creation of digital records uses up our memories and invades our time, and promising platforms seem to solve our needs for finding records and digital resources of any kind. We all know that the outcomes are inadequate without the control on the creation, but we are also aware of the progressive abandonment of our best tools in the management of current records. For this reason, the question at the

[1] Select Committee on Artificial Intelligence of the National Science and Technology Council. "National Artificial Intelligence Research and Development Strategic Plan 2023 Update. A Report", May 2023: VII, https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf. Consulted on 04/06/2023.

beginning of this report is a crucial one: could AI technologies be, more than in the past, the perfect answer to the challenges previously mentioned?

Investigating these issues is a crucial step to continue our mission and live up to our responsibility to protect the rights and memory of people as well as democratic values. This task implies the ability to measure AI technologies, to assess the risks associated with their use, to understand their potential and to develop effective methods for collaboration between archivists and AI specialists.

Based on the above considerations, this report presents the results of the InterPARES Trust AI CU05 study entitled "The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas". In particular, the study aims at assessing whether existing AI technologies can re-establish the archival bond among a multitude of de-contextualized records and to integrate incomplete recordkeeping metadata schemas. In addition, the study aims at identifying archival requirements for AI software, which should be developed according to archival concepts and principles.

1.2 Team members

The study was carried out by the following researchers:

- Stefano Allegrezza (co-chair) (University of Bologna, Italy - Alma Mater Research Institute for Human-Centered Artificial Intelligence - ALMA AI)
- Mariella Guercio (co-chair) (Associazione Nazionale Archivistica Italiana - ANAI)
- Maria Mata Caravaca (International Centre for the Study of the Preservation and Restoration of Cultural Property - ICCROM)
- Lluís-Esteve Casellas Serra (Municipality of Girona, Spain)
- Massimiliano Grandi (Associazione Nazionale Archivistica Italiana - ANAI)
- Bruna La Sorda (Associazione Nazionale Archivistica Italiana - ANAI)
- Francesca Magnoni (North Atlantic Treaty Organization - NATO)
- Samir Musa (Historical Archives of European Union - HAEU)
- Nicola di Matteo (Halifax University, Canada)

1.3 Study approaches

This study has been planned with the aim of supporting the investigation and the archival understanding and knowledge in this area by analysing the most known and promising AI solutions in the specific field of archives and records management and, more specifically, in the creation and/or recovery of the archival relations and original aggregations.

This goal has implied a comprehensive approach, based first on the review of the main platforms available, selected on the following two main criteria: the clear declaration that the document/record management is one of the objectives of the AI based application and the expression of interest for the archival dimension, explicitly stated or easily understandable from the company website. The effort made to limit the number of solutions to be considered and to restrict the survey to market players whose products are relevant to archives and records management (section 2.2) has required time and may have not been free from errors and misinterpretations. Other parameters adopted for the selection have included the analysis of the specific portfolio of the companies, their involvement in the domain of archives and records and their compliance with relevant regulatory frameworks and standards.

Twenty-eight companies were identified and have received an invitation to take part in the survey and answer a very detailed questionnaire. The questionnaire was designed for the systematic collection of the

information necessary for an adequate assessment of the applications intended to support the reconstitution of archival aggregations, as well as metadata enrichment. The questionnaire focus (section 2.3) includes the description of the achievements of the companies, the compliance with the archival regulatory framework, the specific capabilities of the solutions for recordkeeping, the analysis of the AI technologies used and the identification of key performance indicators. Thirteen companies accepted to be contacted and have completed the questionnaire (section 2.3).

1.4 Cooperation with other activities of InterPARES AI-TRUST

There is a connection with study RP03 "Employing AI for Retention & Disposition in Trusted Digital Recordkeeping Repositories (TDRRs)" and a connection with study AA1 "Employing AI for Retention & Disposition in Digital Information and Recordkeeping Systems (DIRS)" leaded by Patricia Franks. The cooperation with study AA1 was aimed at avoiding duplication in contacting AI companies and sharing a common basis for conducting the survey and the interviews with market players involved in Artificial Intelligence with specific focus on records and documents management.

2 METHODOLOGY

2.1 Identification and selection of the AI companies

To achieve the objectives set by Team CU05 it has been crucial to carry out a survey to investigate which AI-based applications currently available on the market can address the needs and requirements of the professional communities of archivists and records managers. An essential stage in this process was the identification of companies that might be interesting for the work of the CU05 team.

A first step was drawing the limits of the pool of the market players prospectively eligible for taking part in the survey. Artificial Intelligence may be used for the analysis and treatment of various kinds of information. In several cases what is on offer has no direct relevance to archives and records management, as – e.g. – there are AI-based products addressing the collection and management of financial data, medical data, images, bitstreams, etc. but not geared at all to catering for any aspect pertinent to archives and records management. Then again, to narrow the number of companies potentially of interest to our study only to those which explicitly state their applications are intended to address archives and records management could have been likely to leave out firms that had developed tools and platforms significant to the objectives of our Team. An acceptable compromise has been stricken by considering – at least for an initial assessment – all those business enterprises including "document management" or – as it is more frequently advertised in corporate websites – "intelligent document processing" among the services they support. In this respect, it is worth pointing out that "intelligent document processing" seems to have become a kind of catchphrase very popular among the market players involved at any level in Artificial Intelligence, and in many cases this label is also used in contexts where the word "document" has nothing to do with what the meaning defined by the InterPARES Project, e.g., when the business activities only deal with the extraction and management of raw data.

In order to select a sizeable number of companies to consider for the survey, our Team relied on two resources:

- direct research on the internet by using various combinations of terms on search engines;

- knowledge accrued by professionals, practices, researchers that belong to the professional communities of archivists and records managers or – at least – are somehow related to them.

As regards to direct research on the internet, the following 9 search strings have been used (8 in English and 1 in German):

1) "artificial intelligence", "records", "documents", "information extraction";
2) "artificial intelligence", "records", "documents", "information extraction", "archives";
3) "artificial intelligence", "records", "documents", "information extraction", "archival bond";
4) "artificial intelligence", "records", "documents", "information extraction", "archives" "archivistics";
5) "artificial intelligence", "records", "document classification", "file plan", "archives";
6) "artificial intelligence", "records management", "document classification", "recordkeeping", "archives";
7) "artificial intelligence", "records management", "categorization", "archives";
8) "artificial intelligence for records management";
9) "künstliche intelligenz", "akten", "dokumenten", "archiven".

All the search terms were connected by an AND operator and Google was the search engine used. This activity – carried out in February 2022 – led to the identification of 21 companies assessed for a possible participation in the survey.

As to the second approach, Team CU05 wish to credit the following professionals who shared their expertise and resources to support our activities:

1) Alan Pelz-Sharpe, founder of Deep Analysis, who shared with CU05 an extremely large database of vendors marketing AI-based products: it goes without saying that this was by far the single resource which has provided the most substantial initial input for our work.
2) Andrew Warland, an information manager based in Melbourne, Australia, who introduced CU05 to two companies whose platforms are of great interest to our research.
3) James Lappin, a doctoral researcher in information management at Loughborough University, UK, who pointed us to some companies potentially relevant to our research and, moreover, found out other resources on the web very useful to enlarge the pool of possible players to be considered for the survey.
4) Jenny Bunn and Paul Young, respectively Head of Archives Research and Digital Archiving Researcher at the National Archives, UK (Jenny Bunn is also a participant in the InterPARES Trust AI project), who helped CU05 to contact one of the companies selected to be included in the survey.

By using the above-mentioned methods and resources, a first assessment stage took place between February 2022 and June 2022, and led to analyse the information concerning ca. 300 companies and to identify an initial group of 100 companies to be further considered for the participation in the survey: the preliminary evaluation was based on the investigation of their respective websites, and the evaluation were based on these criteria:

- statements where the company declares document management as one of the objectives of its AI-based application;
- expressions of interest for any aspect of archives and records management (even if in some cases that is not openly asserted but can only be gleaned from the contents of the website).

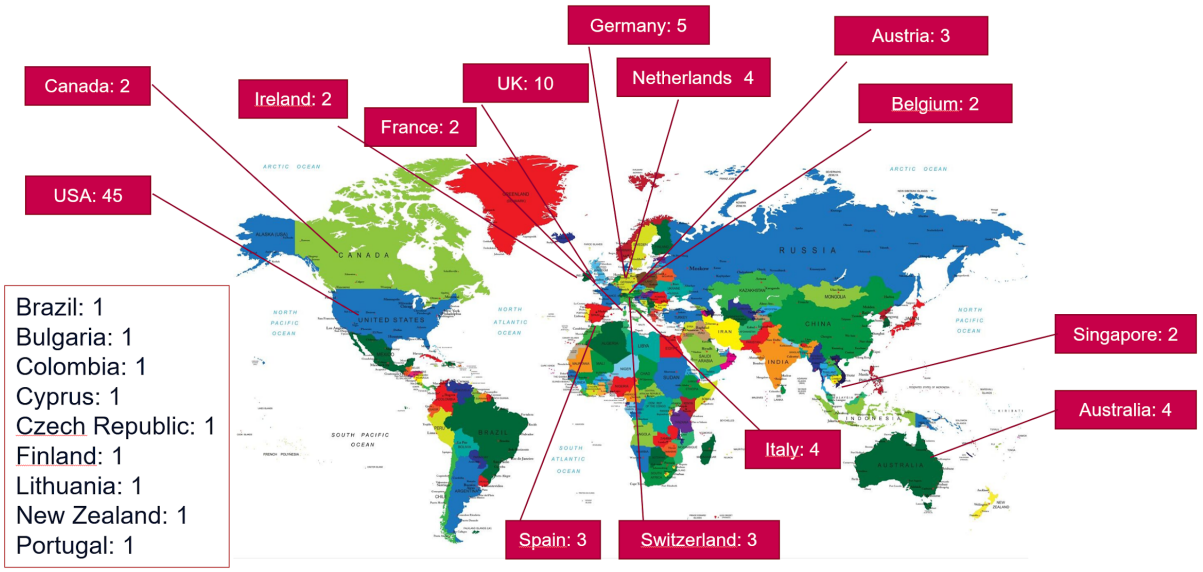

Figure 1. Geographical distribution of the 100 selected companies

The list of 100 companies was further refined in a second assessment round, where a more detailed analysis of the contents of the corporate websites and – where possible – of information gathered through other channels (such as information found in other websites about a company, direct knowledge of the Team CU05 members or references obtained by colleagues or other professionals consulted by us) was carried out. In this round we focused on:

- the specific capabilities and affordances of the AI-based products developed by the companies;
- the specific portfolio of the companies, and in particular their level of involvement with archives and records management;
- compliance with regulatory frameworks and standards relevant to archives and records management;
- the general reputation of the company.

The second assessment stage occurred between June 2022 and August 2022 and led to select a list of 26 companies (later on increased to 28).

2.2 Interaction with AI companies

The research group sent to each of the 28 companies selected an email containing an explanation of the proposal to participate in the research project and stating the objective of the investigation i.e., to understand the AI capabilities for document management / records management. The terms used (document management or records management) has been diversified according to the level of archival expertise possessed by each company.

The research group also attached to the email a letter of invitation containing a presentation of the InterPARES Trust AI Project and its goals, and detailing the research scope of the CU05 Team, namely:

- identifying and reconstituting aggregations of digital records;
- finding appropriate metadata elements to describe such aggregations.

CU05 Team clarified in the letter that the company had been selected after analysing the contents on the website and identifying elements of interest for the purposes of the research.

The invitation letter contained the request to take part in an online interview to investigate the capabilities and affordances of their AI-based products and delve into the underlying technology enabling its operation. CU05 Team also pledged compliance with the protection of the commercial interests of the companies and made it clear that the purpose of the research was to increase the knowledge and understanding of what AI technology may enable archivists and records managers to do, as well as what is possibly needed to improve its performances and mitigate risks.

Invitation letters to each of the selected companies were sent in October 2022, and we got replies from 13 out 28 companies in a span of time from October 2022 to January 2023. All these 13 companies responded by e-mail and agreed to participate in the survey, whose first step was the organization of an online meeting with each of them.

During the online meetings, CU05 Team further detailed the objectives of the research and requested more information on their products, concerning especially both the AI technologies that had been used and specific areas of interest for the research conducted by CU05 Team.

Each company was asked to fill out the questionnaire prepared by CU05 Team.

2.3 Survey Questionnaire

A questionnaire (see Annex 1) was developed to collect more systematically the information provided by the companies whose AI applications are or may be of interest for the archival domain, especially when their products have features useful for the reconstitution of archival aggregations and metadata enrichment.

The questionnaire is divided in **four sections** with open-ended questions.

Section I focuses on the companies' achievements, especially the applications(s) features, development platforms, portfolio, main features and strengths, future developments, aspects to be improved, as well as their compliance with archival and records management standards.

Section II deals with specific capabilities of the applications for recordkeeping, including email management. Questions address the automation of different records management tasks, such as records filing in folders or groups according to a records classification scheme; records appraisal and disposal according to a records retention schedule; extraction of metadata from records; and records indexing to provide information about related aggregations.

Section III analyses the technologies used in the AI applications, such as machine learning models, strategies and techniques; as well as the record or metadata elements the application takes into consideration to make decisions and inferences.

Section IV focuses on audit and key performance indicators to measure the application success rates, and biases.

3 COMPANIES THAT ANSWERED THE SURVEY QUESTIONNAIRE

The 13 companies that accepted to participate to take part in the online meeting and to fill out the questionnaire were:

1) Aluma (UK)
2) Anzyz Technologies AS (Norway)
3) Bis (USA)
4) bizAmica (India)
5) Castlepoint Systems (Australia)
6) Collabware (Canada)
7) Cortical (Austria)
8) Expert.ai (Italy)
9) Grupo adapting (Spain)
10) Iron Mountain (USA)
11) Quest-it (Italy)
12) Read-Coop (Austria)
13) RecordPoint (Australia)

A short presentation for each company follows in alphabetical order below.

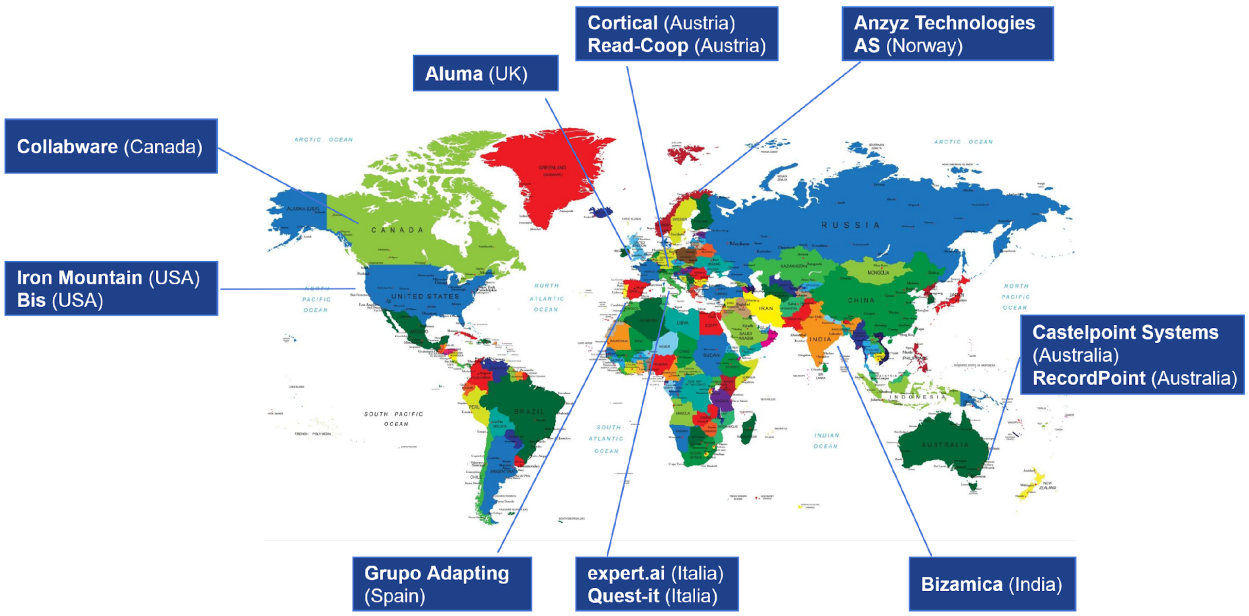

Figure 2. The 13 companies that answered the questionnaire

3.1 Aluma

Corporate website: www.aluma.io

Headquarters: Cambridge, UK; Brooklyn, New York, USA

Introduction: Founded in 2009 - originally as Focal Point Software - by George Harpur and Nidal Husein, Aluma's mission is to develop advanced intelligent document processing technology with a focus on ease of adoption and to blend cutting-edge innovations in data capture, machine learning and information management into tools and technologies that may improve the performance of any digital solution, system, or device. The first cloud service version of Aluma was deployed in 2015. Thanks to their industry partnerships their customers include government bodies, hospitals, and other organisations worldwide.

AI-based product(s) of interest: aluma.io

Short description of the product(s) by the company: aluma.io is a Document Analysis Software Development Kit delivered as a cloud service. Aluma is intended to classify quickly and accurately client documents based on their content, even where the content is highly variable; to extract reliably key data from business documents; to bookmark documents and to make it easy for users to find information and navigate even large document sets.

3.2 Anzyz Technologies AS

Corporate website: www.anzyz.com

Headquarters: Grimstad, Norway

Introduction: Anzyz is an Artificial Intelligence company based in Norway and co-founded by Professor Ole-Christoffer Granmo and Svein Olaf Olsen in 2014. The main developments of Anzyz have been concentrated on search in structured and unstructured text. Their solutions stem from innovations led by Prof. Granmo at the University of Agder, where he is the founding Director of the Centre for Artificial Intelligence Research (CAIR).

AI-based product(s)of interest: 1) CCL™ (Corpus Cube Linguistics); 2) Tsetlin Machine

Short description of the product(s) by the company: 1) CCL™ (Corpus Cube Linguistics) stems from innovations led by Prof. Ole-Christoffer Granmo based on over 10 years of research. The technology is based on Natural Language Processing and is truly unique in a global context. CCL™ combines supervised, unsupervised, and rule-based learning to achieve a very high-level accuracy with significantly less data input needed than traditional Machine Learning. 2) Tsetlin machine is a logical code to train an artificial intelligence in a computer. What is revolutionary about the Tsetlin machine is that it is based on logic as opposed to number-based machine learning. This makes it faster, more cost-effective, and accessible to more users.

3.3 BIS

Corporate website: www.bisok.com

Headquarters: Edmond, Oklahoma, USA

Company description: BIS is a document data integration company. Although BIS's signature product is currently Grooper - launched in 2016, the company itself was established in 1986, at the time when microform scanning, and conversion was their main business. They only use their own staff to develop code and have a large customer base across all industries, in higher education, and in local, state, and federal government.

AI-based product(s)of interest: Grooper

Short description of the product(s) by the company: The first intelligent document processing software and a general-purpose platform for developing Intelligent Document Processing and Data Integration solutions.

3.4 bizAmica

Corporate website: www.bizamica.com

Headquarters: Pune, Maharashtra, India

Corporate website: Founded in 2018, bizAmica is focused on powering businesses with AI for scaling their business: corporate back-office operations cannot be scaled without AI and automation for faster decision making and bizAmica aims at leveraging its AI Platform capabilities to automate document management in such a manner to achieve faster Turnaround Time for decision making, ease of operations, improved efficiency, and significant reduction in processing time.

AI-based product(s)of interest: izDOX Artificial Intelligence Platform

Short description of the product(s) by the company: izDOX AI platform transforms unstructured, semi-structured and structured PDFs, images with printed, handwritten data to meaningful information for quicker decision-making using verification, analysis, and predictions. Higher accuracy, template free solution and multi-lingual capabilities are the product Unique Selling Propositions: izDOX AI uses machine learning, neural networks, natural language processing technologies and their combinations for handling a) auto classification of documents, b) auto extraction of meaningful information, c) hundreds and thousands of variations in the formats, d) handwritten documents and e) multi-lingual documents and at the same time provides higher accuracy.

3.5 Castlepoint Systems

Corporate website: www.castlepoint.systems

Headquarters: Canberra, Australian Capital Territory, Australia

Company description: Founded in 2012, the Castlepoint founders initially established a consulting company, providing expert security, records management, and audit services to the Australian Federal Government and large regulated industry. Later on, Castlepoint has modified its core mission and has focused on developing a way to manage the whole information lifecycle holistically because they believe

security, compliance, and discovery are interdependent. To that end Castlepoint Systems has created the Data Castle paradigm, a solution designed by experts in the field to meet over 900 requirements from laws and Standards: it was built with new AI technology, and architected to be simple, scalable, and secure.

AI-based product(s)of interest: Castlepoint

Short description of the product(s) by the company: Castlepoint is a software solution that manages all the information in an organisation's business systems. It automatically registers every digital record regardless of location or format and uses Artificial Intelligence to classify it against rules and regulations (including secrecy provisions, privacy rules, and Records Authorities for example) and apply appropriate lifecycle controls. It acts as a single interface to find, relate, manage, and audit every record in an organisation's network, no matter what system it is stored in, and it does this without any impact on existing systems or users, and without complex rules engines. This allows governance teams to finally have a complete view across the whole environment, and to apply security, discovery, and compliance processes to every single system from a single interface.

3.6 Collabware

Corporate website: www.collabware.com

Headquarters: Vancouver, British Columbia, Canada; Washington, District of Columbia, USA

Company description: Founded in 2010, the core mission of Collabware is providing Intelligent Enterprise Content Management solutions and services to free organizations from information chaos. It cooperates with ARMA (Association of Records Managers and Administrators) and AIIM (Association for Intelligent Information Management) and is one of the partners of InterPARES Trust AI.

AI-based product(s)of interest: 1) Collabware CLM; 2) Collabspace; 3) Collabmail

Short description of the product(s) by the company: 1) Collabware CLM is an on-premise solution for Sharepoint to help automate records management; 2) Collabspace is a cloud-based information governance suite which can connect to Office 365 and several other content sources: it includes three different products a) Collabspace Archive, b) Collabspace Discovery, c) Collabspace Continuum; 3) Collabmail is an Outlook attachment that allows for the easy filing of emails to Sharepoint Online.

3.7 Cortical

Corporate website: www.cortical.io

Headquarters: Wien, Austria, Europe; New York, New York, USA; San Francisco, California, USA

Company description: With more than 10 years expertise in implementing intelligent document processing solutions in the enterprise, Cortical.io's mission is to deliver AI-based solutions that streamline the extraction, classification, review, and analysis of information hidden in unstructured text while providing short time to value. They accomplish this through a novel, meaning-based approach to natural language understanding that aims at solving many critical challenges of text processing in a business context and the problems of language ambiguity and variability across many use cases and verticals.

AI-based product(s)of interest: SemanticPro

Short description of the product(s) by the company: SemanticPro automatically and accurately extracts, searches, compares, classifies, and routes large volumes of unstructured documents like contracts, leases, insurance policies, emails with attachments, and message streams at scale. It combines high accuracy and speed in processing complex content where language needs to be interpreted, enabling the automation of traditionally labour-intensive workflows.

3.8 expert.ai

Corporate website: www.expert.ai

Headquarters: Modena, Italy

Company description: expert.ai's core mission is to help organizations turn language into data to make better decisions. They are a team of AI experts using the full range of natural language technologies to analyse data, understand it to improve knowledge extraction and accelerate intelligent automation. With 250 customers globally, expert.ai has offices in Europe and North America.

AI-based product(s) of interest: expert.ai Hybrid Natural Language Platform

Short description of the product(s) by the company: The expert.ai hybrid natural language platform provides a deep understanding of language, from complex documents (e.g., contracts, emails, reports, etc.) to social media messages, and turns it into knowledge and insight. This makes for faster, better decisions without all the manual, time-consuming work. Key capabilities of the platform include: 1) Language Data Tools; 2) Intelligent Document Processing; 3) Hybrid AI (i.e., it adds human-in-the-loop (HITL) feedback to fine-tune models and capture subject matter expertise); 4) Domain Knowledge Models; 5) Natural Language Processing Ecosystem.

3.9 Grupo Adapting

Corporate website: www.adapting.com

Headquarters: Valencia, Spain; Barranquilla, Colombia

Company description: Established in 1999, Grupo Adapting's core mission is to develop professional record management systems with a particular focus on the Spanish and Latin American market: the main office in Latin America is in Barranquilla, Colombia. Grupo Adapting pledges compliance with several standards relevant to records management - such as ISO-15489, ISO-16175, ISO 19005-1:2005, Moreq, ICA, DoD 5015, ISO 30301 - and is committed to designing AI-based solutions for records management

AI-based product(s)of interest: 1) Abox-ECM; 2) Cbox-Cloud

Short description of the product(s) by the company:

1) Abox-ECM is a professional document management solution that captures, stores, processes and distributes documents. Thanks to AI Abox-ECM can automate document classification. Abox-ECM is available in for different models: Abox Entry, Abox Plus, Abox Archive and Abox Elite; 2) Cbox-Cloud is the cloud version of Abox-ECM.

3.10 Iron Mountain

Corporate website: www.ironmountain.com

Headquarters: Boston, Massachusetts, USA

Company description: Iron Mountain is one the largest companies among those whose core business and mission is the delivery of records management services: it was founded in 1951 and has established itself as a major player in its specific industrial vertical sector. Iron Mountain is a brand well-known by the professional communities of archivists and records managers and is one of the partners of InterPARES Trust AI.

AI-based product(s)of interest: InSight

Short description of the product(s) by the company: AI-based Intelligent Document Processing studio to process the data, and content service platform to host, review and search the data.

3.11 Quest-it

Corporate website: www.quest-it.com

Headquarters: Siena, Italy

Company description: Founded in 2007, Quest-its core business is to develop AI solutions based Natural Language Processing to support and improve information retrieval, information extraction and intelligent question and answer platforms.

AI-based product(s)of interest: Detecto Evo

Short description of the product(s) by the company: Detecto Evo is a document management platform to help users to digitize, classify and store the data of their business documents: it can intelligently recognize the type of a document, identify the kind of information contained in the text (amounts, letters, punctuation, reasons, date, signature, etc.), carry out a semantic analysis of the words, move the document in the appropriate section of the repository and classify the document by examining its content and the position of the text in the document.

3.12 Read-Coop

Corporate website: www.readcoop.eu

Headquarters: Innsbruck, Austria

Company description: Read-Coop is a European Cooperative Society with more than 100 Members globally. Read-Coop SCE (Societas Cooperativa Europaea) with limited liability was established on 1 July

2019, to sustain and further develop the Transkribus platform. Transkribus was developed within the Horizon 2020 “READ” European Union project by a consortium of leading research groups from all over Europe, headed by the University of Innsbruck.

AI-based product(s)of interest: Transkribus

Short description of the product(s) by the company: Transkribus is an AI-powered platform for text recognition, transcription and searching of historical documents – from any place, any time, and in any language. Transkribus enables the automatic recognition of text, layout, and structure in documents with the power of AI: users can train their own AI models that fit their specific documents. Transkribus also lets users enrich their material with metadata.

3.13 RecordPoint

Corporate website: www.recordpoint.com

Headquarters: Sydney, New South Wales, Australia

Company description: RecordPoint was founded in 2009 by Elon Aizenstros and Anthony Woodward on the idea of democratizing access to data and increasing data trust, with the mission of building technology that helps organizations be more trusted: they want to tackle the problem that too often data is stored in silos across a patchwork of legacy platforms - which restricts its access and use - and is kept for far too long. RecordPoint also cooperates with the UK National Archives and one of RecordPoint’s founders, Anthony Woodward, contributed to the MoReq 2010 standard.

AI-based product(s)of interest: Records365

Short description of the product(s) by the company: Records365 is an in-place records management platform that can ingest data from any digital content source. The RecordPoint Data Trust platform helps highly regulated organizations manage records and data throughout their lifecycle, regardless of system. It’s a manage-in-place platform that can ingest data from any content source. RecordPoint’s capabilities span six core areas, which are the essential building blocks for solid data governance - data inventory, data categorization, records management, data privacy, data minimization, and data migration. The platform classifies records to determine their lifetime and will dispose them when disposal is due. Automated classification comes in two flavours: expert system (rules-based) classification that uses record metadata, and machine learning classification, based on record text. Records365 also enriches records by extracting personal information, named entities, and other signals from text and metadata content, allowing for sophisticated federated search and reporting. Legal holds and physical records are also supported.

4 ANALYSIS OF THE QUESTIONNAIRE ANSWERS

4.1 Outlines of the AI companies

All the companies interviewed have developed solutions based on artificial intelligence technologies to manage, index and classify structured, semi-structured and unstructured data through automatic learning techniques and automatic data extraction.

The main purposes stated by the interviewed companies relate to specific functions. Although it is possible to identify macro – functions, it is important to consider that each company is carrying out research into a specific solution according to different activities and purposes. See list of macro-functions and specific activities below:

1) Automatic document management

- Interpreting and classifying information from complex documents;
- Archiving data;
- Implementing a "No tagging" solution for datasets, for both structured and unstructured data;
- Reducing manual review cycles subject to errors;
- Reducing risks, error rates and revision times;
- Building and perfecting personalized classifiers;
- Monitoring "performance with integrated dashboard with drill-down capabilities";
- Allowing the governance teams to have a complete view across the whole environment;
- Applying security, discovery and compliance processes to every single system from a single interface;
- Developing an easier distribution (web client, better cloud/containerization support, etc.).

2) Automatic extraction of information

- Automatic extracting of significant information;
- Managing variations in documents;
- Enriching contents;
- Creating connections of entities;
- Elaborating a tailor-made dashboard for text analysis e.g., research and investigation on text data including metadata; in this respect, some learning algorithms automatically include incorrect and synonymous words in the user's search, leaving no data left behind;
- Managing multilingual documents in any written text format including errors and increasing the ability to learn specific domain terminology;
- Creating support tools for the generation of statistics and statistical categories;
- Generating automated and deep insights into data sets without manual tagging, and without the use of predefined concepts for deeper data insights;
- Building high-quality and interpretable models that can help to understand data;
- Automating traditionally high-work intensity workflows.

3) Management of hand-written documents

- Recognizing handwritten text as well as layouts and tables, extracting information, labelling;
- Improving the accessibility of the archival material, through the training of specific recognition models, the ability of carrying out full-text searches and the capability of reading historical handwriting.

4) Research capabilities

- Developing solutions for research consisting “of a user-friendly dashboard with an API and background AI trained data”;
- Performing advanced semantic research;
- Improving research capabilities.

A single company said that their algorithm falls within the “category of ‘green AI’ due to its significantly reduced need for computation and human intervention (no manual tagging)”.

Companies most involved in records management have also included, among their purposes, functions specifically focused on the documentary environment and its life cycle such as:

1) Automated classification

- Providing automated registration, classification, and retention/disposition management for any business system, through automated and scalable record management;
- Automatically registering every digital record regardless of location or format, by using A.I. to classify it against rules and regulations and to apply lifecycle controls;
- Automatically classifying all contents against security markers, secrecy provisions, record retention rules, inquires, investigations, and other mandatory policies to meet governance and compliance requirements (IT Security, Internal Audit, FOI/Legal Teams and Records Manager) by combining their classification taxonomies;
- Automatically classifying documents based on logical content rules created by the customer and, besides that, tracing and automatically pushing the records through each phase of their conservation, including time-based and event-based retention;
- Using “an expert system (rules-based) classification that uses record metadata, and machine learning classification, based on record text”;
- Creating an inventory of all the structured and unstructured data of an organization “as a part of a continuous inventory process across the data lifecycle”.

2) Records and data management

- Guaranteeing a quick configuration (on-the-fly machine learning and reusable configuration models), simple setup and easy deployment through cloud service;
- Assisting highly regulated organizations in managing records and data during their life cycle, regardless of the system, through a manage-in-place platform that can ingest data from any origin of content;
- Managing physical records;
- Tracking physical and electronic records.

3) Migration

- Identifying high and low value data before migration;
- Appraising legacy content repositories (such as file shares) and identifying various data trends and insights with a view to preparing a migration.

4) Information content management

- Managing all the content in thousands of systems through a single interface;
- Providing end-to-end intelligence for the entire life cycle of information;
- Using the processing of natural language to understand what each document, email or database row is about, by extracting key phrases and named entities.

5) Use of metadata

- Using the metadata of records in their source of content to keep track of all changes through a complete audit system;
- Using the processing of natural language to understand what each document, email or database row is about, by extracting key phrases and named entities;
- Looking for records through the construction of advanced queries that can exploit almost all metadata fields for record management purposes;
- Enriching records by extracting personal information, nominated entities and other signals from the text content and metadata, allowing for sophisticated federated search and reporting.

6) Risks management

- Preparing automatically flags and alerts for high risk or sensitive data, based on operational risks;
- Capturing all events across the enterprise for audit, security, and integrity, and ensuring that even the deleted items have a permanent record activity.
- Providing means for the systematic review of records in order to perform a complete and irreversible digital destruction;
- Providing executive level dashboard of information risk;
- Managing privacy information and delivering scalable solutions for data discovery, data privacy, data categorization and data minimization;
- Reducing the risk of unauthorized access.

7) Legal holds management

- Supporting legal holds;
- Responding to freedom of information (FOI) and the right to be forgotten requests.

The interviewed companies work more with other business firms, but almost all of them have relations with universities, archives, and governments, state, local and international government agencies, health care, media, and editor companies. The main work areas are: company systems for research and management applications, various sectors such as banking and finance, insurance, production, logistics, commercial operations, funding, sales and legal customer service (energy, financial services, public services), human resources (CV evaluation, interviews powered by artificial intelligence), digitization of documents, intelligent research in multimedia repositories, legal sector, public services, energy, and financial services.

In order to improve operational functionalities, companies indicated as future steps of their research and development activities the following tasks:

1) Records management

- Improving solutions to make indexing and classification faster and faster, while scaling "*across larger and larger networks managing terabyte and petabyte of data*";
- Establishing aggregates i.e., being able to define parent-child relationships between records, so as to allow users to accommodate case files, folder structure, and other scenario-dependent records management scenarios;
- Examining new ways to combine the classification of the text with the rules and to offer new classification taxonomies in addition to record sentencing;
- Developing clustering applications to identify groups or record without being trained on a predefined taxonomy;
- Developing unsupervised learning methods based on clustering and automatic document type assignment;
- Correctly capturing documents and metadata;

- Creating an intelligent record assistant with AI that may be able to answer complex questions about records by studying data and metadata;
- Searching for new automatic learning algorithms for the separation of documents;
- Expanding the range of documents and types of text that technology can include correctly;
- Searching for new automatic learning algorithms for the separation of documents.

2) AI bases improvement

- Improving the "automatization of pre-processing and producing AI bases";
- Training at the same time several bases of artificial intelligence.

3) Content enrichment

- Improving the quality of the analysis and the depth of understanding;
- Fully integrating content enrichment like sentiment analysis and entity extraction;
- Developing tools for automatic audio transcription and the identification of speakers;
- "Bringing some new AI capabilities into the platform, including question answering, text synthesis, translation, and image recognition";
- Building enhanced image processing for low-quality documents.

4) General purposes

- Adding "more powerful and easier tools (…) for a faster and more effective integration of technology itself, covering more and more complex business cases";
- Identifying new commercial possibilities to exploit the power of digitization to help businesses and companies to leverage successfully Big Data;
- Ensuring "scalability and security so as to stay on top of exponentially growing data volumes and ever-more-sophisticated cyber-attacks";
- Enabling an empathic discussion with users using innovative technique such as emotion detection, Named Entity Recognition, dialogue management;
- Using Learning Intelligent Systems in order to support hearing-impaired people;
- Improving usability, performance, speed, reliability.

The richness of answers testifies differentiated approaches, aiming at supporting innovative abilities in the extraction, management, and restitution of information content.

The variety of creative solutions listed in the portfolios could be the consequence of the complex tasks needed to respect the peculiarities of the archival requirements or reflects the nature of dynamic technologies dominated by an ongoing process of evolution and transformation.

4.2 Involvement with records management and archives

Among the companies interviewed, only five said they have not had any relations with archives and records management so far. All the others liaised with national, territorial and health archives for specific research purposes, such as the automatization of solutions for mail archiving, the creation of databases for patient health records, the handling of handwritten or multilingual documents, the extraction of metadata elements used to support a better search experience, solutions for searching and browsing digitalized collection and the improved accessibility of archival materials.

Companies mostly involved in archive and records management underlined their capabilities in different processes such as document classification, indexing, managing the whole life cycle of documents and records, including the accession to archives.

Some solutions are designed for automating records management. A company wrote that "using metadata from the records at their content source, we use WORM[2] storage and the data lake to keep track of all changes via a thorough audit system, and can automatically classify documents based on logical content rules created by the customer" and that they "can automatically track and push records through each phase of their retention, including time-based and event-based retentions; we provide means for systematic review of records, and then perform complete and irreversible digital destruction".

The applications declare that they comply with the following standards (or have been designed to support them):

- ISO 15489 (Records management);
- ISO 16175 (Information and documentation — Processes and functional requirements for software for managing records);
- ISO 23081-1:2017 (Information and documentation — Records management processes - Metadata for records);
- ISO 30301:2019 (Information and documentation — Management systems for records — Requirements);
- ISO/IEC 27001 (Information security management systems);
- MoReq 2010 (Modular Requirements for records systems);
- UNI-EN ISO 9001 (Quality management systems);
- DoD 5015.02-STD (Design Criteria Standards for Electronic Records Management Software Applications);
- FINRA-SEC 17A-4 (Electronic Recordkeeping Requirements for Broker-Dealers, Security-Based Swap Dealers, and Major Security-Based Swap Participants);
- US SOC2 Type 2 (System and Organization Controls);
- FedRAMP Medium authorization ("The FedRAMP PMO fields a number of questions about impact levels and the security categorization of cloud services. Federal Information Processing Standard (FIPS) 199 provides the standards for categorizing information and information systems, which is the process CSPs use to ensure their services meet the minimum security requirements for the data processed, stored, and transmitted on them. The security categories are based on the potential impact that certain events would have on an organization's ability to accomplish its assigned mission, protect its assets, fulfill its legal responsibilities, maintain its day-to-day functions, and protect individuals")[3];
- GSA Advantage;
- NARA M-19-21 ("M-19-21 is a memorandum issued by NARA on June 28, 2019. A consolidation of the previous M-12-18 directive with some additional requirements, the purpose of this directive is to help the government transition fully to electronic records for increased efficiency, accuracy, and improved storage[4]");
- CAN/CGSB-72.34-2017 (Electronic records as documentary evidence)[5];
- DCAM (Data management capability and Assessment Model);
- CCPA (California Consumer Privacy Act);

---

[2] WORM stands for "Write Once Read Many".

[3] https://www.fedramp.gov/understanding-baselines-and-impact-levels. Consulted on 04/06/2023.

[4] https://info.aiim.org/aiim-blog/directive-m-19-21-what-it-is-and-how-to-achieve-compliance. Consulted on 04/06/2023; as to the link to the Memorandum see https://www.archives.gov/files/records-mgmt/policy/m-19-21-transition-to-federal-records.pdf. Consulted on 04/06/2023.

[5] https://publications.gc.ca/site/eng/9.839939/publication.html.

- GDPR (General Data Protection Regulation).

4.3 Capabilities relevant for records management and archives

4.3.1 Records organization

4.3.1.1 Classification

Classification is the act of linking records to their business context (to the business being documented). This can be accomplished by associating records with categories in a business classification scheme, thus records are linked at an appropriate level or class (for example, to a function, activity or work process).[6]

The automatic classification of records, which uses the support of a classification scheme, is offered by almost all the companies that replied to the questionnaires (10 out 13), including those not specifically involved with archives and records management. Among the companies whose platforms do not feature automatic classification, one mainly deals with handwritten text recognition and the other two focus on indexation and extraction of metadata elements. Some of these affirm their applications could be modelled or trained to categorise records.

Three companies familiar with the principles of archives and records management replied that their platforms analyse the metadata elements available both in the records and aggregations of which they are part, in particular:

- one pointed out that its AI application classifies records "according to document type and case-folder specifications";
- the second one replied they classify documents and aggregations "based on their content and context, against function-based records classification schemes";
- the platform developed by the third one initially tries to classify the documents and aggregations by analysing "any metadata fields that are found or inferred on records". In case the available metadata should prove to be insufficient for classification, then their software "will use a machine learning model (if configured) to classify the record based on its text". The company also pointed out that the "taxonomy, rules and ML model are fully customizable by each customer and can be used to reflect any classification scheme and administrative processes required", although – as to administrative processes – at present "only disposal workflows are currently supported based on the classification".

Six other companies (two of which are familiar with archives and records management) concisely answered that their platforms are able to categorise documents according to a records classification scheme: one of them also added that they have developed "a classifier that is able to classify documents and images".

The platform of the thirteenth and last company can be trained to recognize document types and apply any kind of classification scheme - more specifically, they say their platform is able to "process files and records (with suitable connectors to the source systems) and generate labels and tags belonging to any record classification scheme (taxonomy or term ontology)". In accordance with the general approach adopted by this company to develop its products, the platform can be trained by the users themselves, who can feed the AI application with specific sets of data: based on the input, the platform builds a knowledge graph that will be progressively refined and used to classify records and aggregations.

---

[6] International Organization for Standardization, BS ISO 15489-1:2016: Information and Documentation – Records Management, Part 1: Concepts and principles, The British Standards Institution 2016: 17.

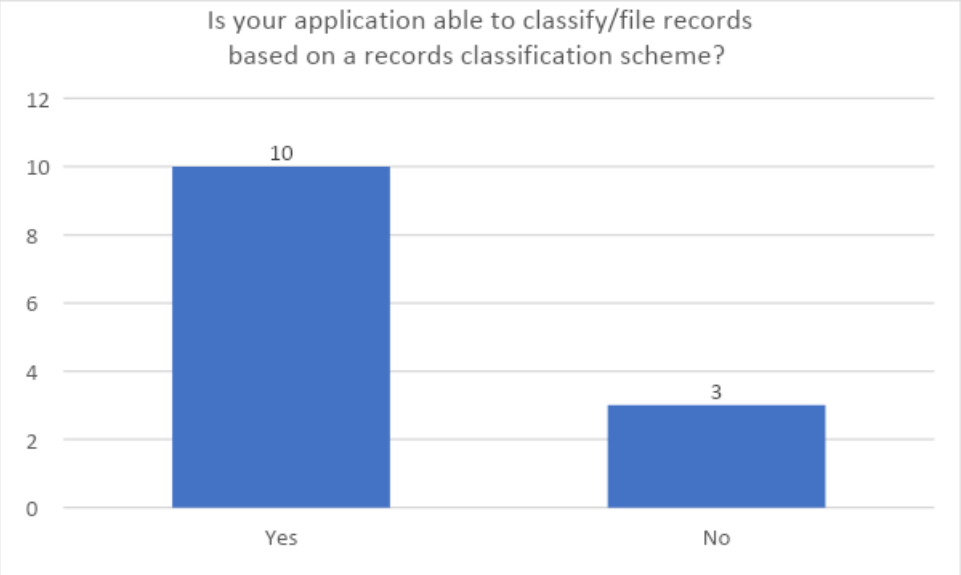

Figure 3. Classification capabilities

4.3.1.2 Aggregation

While classification deals with providing context to a record and establishing relationships between a record and the activity for which it was created, aggregation is the act of grouping together related records[7] that, if combined, can constitute different archival units, such as files or series.[8]

The concept of aggregation is equated to filing. In the Italian archival tradition, while "classification guides records sedimentation in an orderly and consistent manner; filing aggregates all the records produced by the same activity or administrative process into archival units. Therefore, classes and files are separate but interrelated entities of the same structure [the records classification scheme, integrated with the file plan]. Classes represent the functions and activities attributed to a records creator through regulation. They form an abstract structure in which, generally at the last classification level, files are created. Records are preferably placed into files or are logically linked to them."[9]

As per the questionnaire, 10 out 13 of the interviewed companies stated that their applications are able to file records in their related folders, case-files or groups in an automatic or semi-automatic way, although with limitations:

- "There are some automatizations with the document type (associated with a folder or case-folder)";
- "Records are aggregated to a parent that users can specify per content source, and which is reflective of the structure of the content source" (e.g. an email inbox, a folder on a file share)";
- "It can be used to automatically (or, in some cases, semi-automatically, with a final user validation in the case of a human-in-the-loop workflow) generate metadata labels, tags and other information

[7] "ICA-Req: Principles and Functional Requirements for Records in Electronic Office Environments, Module 2: Guidelines and Functional Requirements for Electronic Records Management Systems", 2008: 26.

[8] DLM Forum Foundation, "MoReq2010: Modular Requirements for Records Systems", Volume 1, "Core Services & Plug-in Modules", Version 1.1, 2011.

[9] Mata Caravaca, María, "Policies and Requirements for Archival Sedimentation in a Hybrid Records Management Environment: A Critical Analysis of International Writings", PhD Thesis, Sapienza Università di Roma, 2017: 38.

on the base of the document content." The generation of labels and tags may follow "one or more classification or labelling schemes".

Among the three companies that do not provide automatic filing, one company declared that its application "is capable of automatically categorising records based on their metadata and content rules created by users, but it does not file items to folders at this time." They are in the process of developing this function.

Additional questions were made in relation to the creation of aggregations, such as the possibility that applications make inferences about which records belong or might belong to the same group or business process (e.g. same case-file, subject-file, series, fonds). Answers were positive in the range of 9 out 13 as well, indicating that inferences are made as follows:

- "Based on content and/or context";
- "Case-folder might be found using IA models in a trained system";
- "Only with user-created content rules to specify the categories and requirements for categorization. Once these content rules are created and enabled, the system can make these decisions on newly added records";
- "If there is metadata to represent those processes (e.g. a case file number)";
- "The tools can output aggregated extractions exposing specific relations between extracted entities".

One company replied that its application could not make inference directly, but could provide data which could be used for this purpose.

When companies were asked if their application was able to make inferences about the organisation or person that filed the records, even when relevant metadata elements for their identification were missing, the number of positive answers decreased to a range of 7 to 13. Those answering positively, wrote:

- "Getting the right case-folder and document type, you can establish sender and receiver of the record";
- "If the person, or organisation is clearly marked on the document e.g. with a stamp this should be possible";
- "If the involved entities are stated in the content of the document and the type of relation linking them is linguistically expressed, the technology is able to make such inferences";
- "If the data somewhere exists in the records, then without metadata identification is possible";
- "It will only identify organisations or persons referred to in the record text. It will not infer the role of that entity in the provenance of the record";
- "Yes, based on content and/or context";
- "Possibly, but through classification".

Those responding in a negative way, expressed that inference could not be made directly without metadata, for example: "Our entity extraction feature can identify people, places, and things from the extracted text of a document, but if metadata is missing there is no framework as of yet to automatically make inferences about it".

In summary, developers of applications featuring automatic classification capabilities assert that applications are also able to file records or logically link records to their related aggregation, as well as make inferences to achieve this goal. They also proclaim that this automation may follow classification or labelling schemes, and it is based on how the applications work or how they are trained, that is, according to record type, case-file, metadata, record content and context.

These assertions are perhaps too optimistic. Both classification and filing/aggregation have different connotations depending on the archival tradition or the different records management standards. These

may distinguish between both activities, or may not consider distinction, or may confuse them. This is, in some way, reflected in the answers to the questionnaire, where is not always easy to understand if applications categorise and/or set aside records based on established classes and/or files, and to what extent they proceed following a records classification scheme and/or a file plan, or just an established list of labels or tags. It should be specified that non all the companies were familiar with some of the archival terminology used in the questionnaire, even if the sense of the questions were explained in the introductory interviews held with them. This could have caused a lack of understanding of the questions and not clear answers. For this reason, the effective and efficient application of these automated platforms to specific records management needs would have to be analysed more in depth in the second part of this project, which will focus on testing the applications with records series belonging to institutions participating as case-studies.

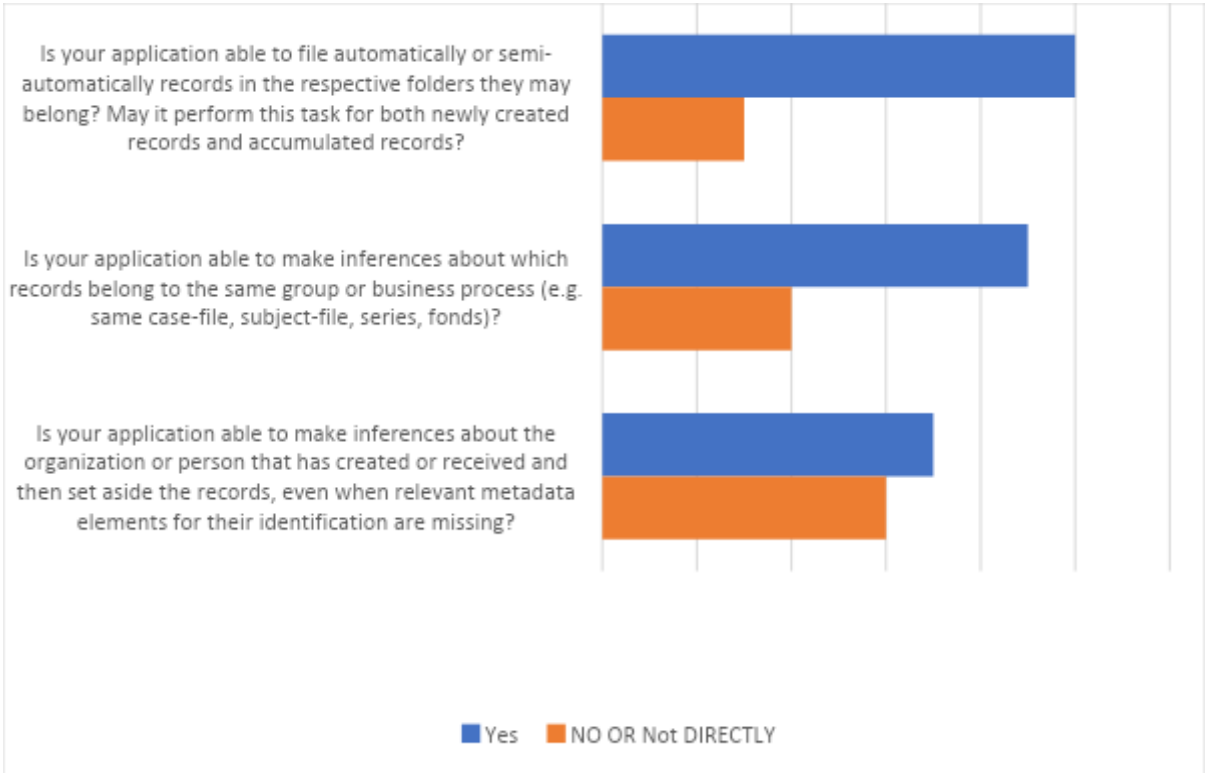

Figure 4. Aggregation capabilities

4.3.1.3 Reconstitution of the archival bond

Could existing AI technologies re-establish the archival bond of non-aggregated, unarranged, or de-contextualized records, both in the current and semi-current phases of their lifecycle, basically to ensure an accurate appraisal and guarantee proper transfer procedures?

The companies were asked this question, specifically if their applications could re-constitute archival aggregations that had been lost. Five out 13 answered positively. They basically extract data from records content or from metadata (including the classification schemes) to propose aggregations or relations among records, for example:

- "Our system contains clustering and 'find similar' technology which is able to propose aggregations and/or can be used to extract data that can be used for this purpose".

- “Yes, in most circumstances, but this is often contingent on the metadata available from the content source”.
- “Any lost data can be regenerated if source-digital documents are available. The source digital documents can be used to extract the lost data”.
- “This re-construction can be performed […], but limited to the cases where the aggregation logics can be extracted from the document text content”.
- “Probably - we can perform classification, and what this describes is ultimately a classification problem”.

The other eight companies replied in a negative way, adding that this is something to be done only manually, or that they could “provide restored records to users in the case of data loss at the content source, but cannot reconstitute aggregations”.

The interviewed companies were also asked if their applications were able to index records in order to provide information about related links or aggregations among records. Six out of 13 answered positively:

- “We capture all contextual metadata including relationships between items and aggregations”;
- “All data in the system is always indexed”;
- “The technology does not natively work with record management, but can extract information from the text by understanding the content meaning and make it available to downstream record management systems which can index this information to create links, aggregations, etc.”

The other companies replied in an undetermined or negative way:

- “Maybe, but maybe only in a very limited way”;
- “Not explicitly, but if there is metadata available on the aggregation or record in the content source that will permit this then it is possible”;
- “This feature is in development at this time, but the application is able to identify similar emails based on indexing”.

Thus, reconstituting records aggregations is not the main focus of companies developing AI technologies. It is a difficult task if contextual data is lost. This will require a more evolved technology.

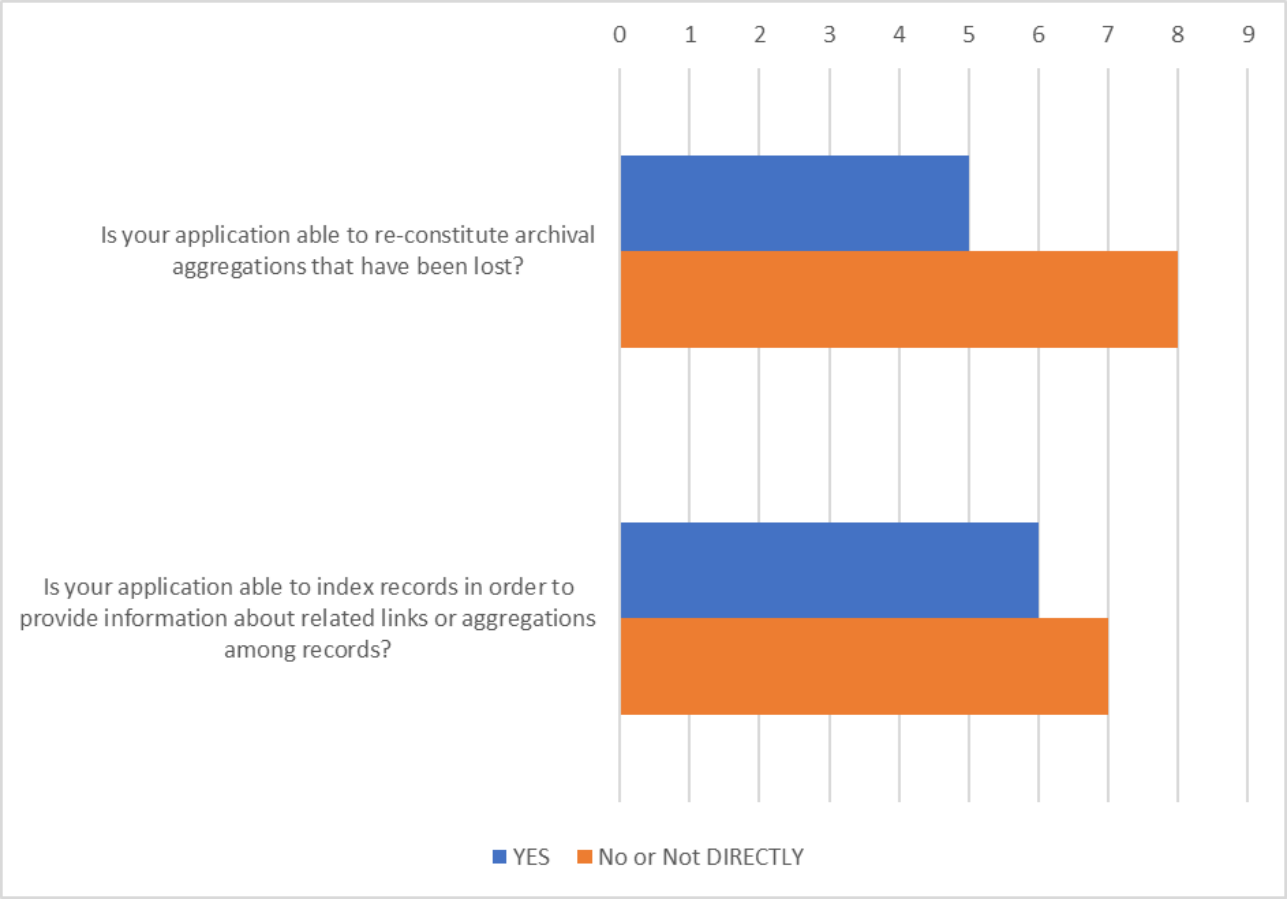

Figure 5. Reconstitution of archival bond capabilities

4.3.2 Extraction and indexation of metadata

Metadata is defined in the technical literature simply as "data on (other) data, information that describes a set of data. Metadata is used for multiple purposes: search, management and localization, selection, interoperability.[10]

As Giovanni Michetti (2014) has pointed out: "The management of the huge amount of documents produced in a digital environment requires the adoption of a robust metadata system to support the processes of creation, processing and storage of documents […]. Metadata documents the production context of document objects and therefore must be identified according to the context itself, as well as the specific purposes that justify its creation and use."[11]

An index is defined as a list of keywords associated with a record, used especially as an aid in searching for information. The main purpose of a Record Indexing System "is to facilitate the indexing of records in the recordkeeping system, through assignment of access points to each record using a controlled recordkeeping vocabulary, for the purpose of facilitating effective and efficient discovery and retrieval of records in the recordkeeping system."[12]

---

[10] M. Guercio, *Archivistica informatica,* Roma, 2010: 189.

[11] G. Michetti, "Gli standard di gestione documentale", in *Archivistica. Teorie, metodi e pratiche*, 2014: 269.

[12] Cf. T. Eastwood, H. Hofman and R. Preston, "Chain of Preservation Model Narrative" in "*InterPARES 2 Project Book. Part Five: Modeling Digital Records Creation, Maintenance and Preservation*". Rome, Italy; ANAI 2008: 202.

Developing methods and tools for automatic normalized metadata capture is central to the integrity and authenticity of digital sources. AI solutions can increase metadata enrichment and content indexing to further expand the links between concepts in ontologies and provide even easier and faster access to content of interest.

All the companies have developed solutions to extract metadata, by using e.g., parsers to extract required entities or external automatic OCR applications to fill the metadata fields of the document to describe it or to generate metadata elements from the content, where OCR is possible. A company answered that "Handwritten text can sometimes be effectively processed with software-based OCR tools, but not commonly" and that usually they "that handwritten documents need to be OCRd by a hardware scanner which are much more sophisticated for this purpose". Some companies noticed that the quality of the text extracted through OCR processing is questionable and is expensive to run.

By reviewing in more detail the replies of some of the companies that have been surveyed:

- a company declared they "can extract information from scanned documents and this information may be used also to create records";
- another company wrote they "can map metadata from most systems we connect to, can extract from the text of the document itself, and can perform ICR on handwritten text using a built-in Azure CV integration";
- a third company declared their platform "scrapes extrinsic metadata from the content source and sometimes from intrinsic metadata from inside the record files" and that it "also has the capability to add metadata defined by external systems using pattern matching rules. Some metadata is derived by AI technologies from the record text". They have also added that their products "has a 'signaling' process in the enrichment pipeline that calculates new metadata based on the less refined metadata the system has harvested or mined";
- a fourth company replied that their application "does not natively work on source handwritten documents, but can be integrated with existing tools that can perform the job (OCR, ICR, etc.) thanks to a number of existing connectors, and that can work upstream. The output of these tools can be plugged to the technology components, so they can use the generated text, and possibly identify the document structure where needed to properly generate metadata, tags, labels, etc. for the business purpose."
- two companies wrote they can use Natural Language Understanding (NLU) for the extraction and indexation of metadata. NLU can use precise and detailed indexing and semantic tagging to make content even more accessible and usable by humans. By enabling the creation of more links between topics and concepts, NLU lets the full value of information emerge, making it easier for researchers to retrieve relevant content and identify new connections between multiple elements, topics, and metadata.

4.3.3 Appraisal and retention

Appraisal has been defined by InterPARES 2 as "the process of assessing the value of records for the purpose of determining the length and conditions of their preservation."[13] Appraisal is therefore closely related to retention, as a decision about the length of retention always depends to some extent on an appraisal process, that can be performed at various stages of the document life-cycle and is a process which may also be repeated.

[13] Cf. "Terminology Cross-domain Task Force. Glossary". In *InterPARES 2 Project Book*, Rome, ANAI 2008: 772.

In principle, AI-based applications might help humans to assess whether a document is likely to be a record or not, or even provide more in-depth analyses by evaluating features and parameters to gauge the archival value of records according to a pre-established set of criteria. As in many other aspects, AI might be useful particularly when huge numbers of documents are involved.

We asked the companies whether their products: 1) are able to carry out any appraisal activity and identify records with archival value; and 2) can identify records to be disposed of, based on records retention schedules.

The first question was rather a general statement, as “appraisal” is a word often not understood in the same way as archivists and records managers do by people who do not belong to their professional communities: a hint at the identification of the records of archival value has been added to give the companies a broad definition of the purpose of the appraisal activity.

The second question was instead somewhat specific, as we have tried to pin down the most common outcome many customers of our responders may seek when they implement retention schedules i.e., clarity about whether and when they can carry out the permissible destruction of their records.

As to the first question about appraisal activities (no. 11 of the questionnaire):

- Six companies replied that their products do not perform any appraisal activity, although one of these specified that their platform can receive and elaborate a “human input” to appraise documents i.e., humans can directly specify information concerning the appraisal of the documents.
- One company replied that its platform in its original configuration (i.e., the “vanilla” platform) does not possess such a capability, but they can build or connect with a separate module and customize their product so as to enable it to execute some appraisal activities.
- Six companies replied that their products can carry out appraisal activity, although three out of six have not further elaborated on their reply.

As to the other three companies, their products can carry out appraisal activities:

- Two linked appraisal to classification; one replied that “We could do this using our classification techniques” and the other one said “Valuable records are usually identified and tagged using the standard classification scheme”.
- One company wrote that through content analysis its product may implement “business logics that can enable prioritization of the structured information extracted” and through this process assign scores by means of which “send alerts” to human operators when given threshold values are attained.

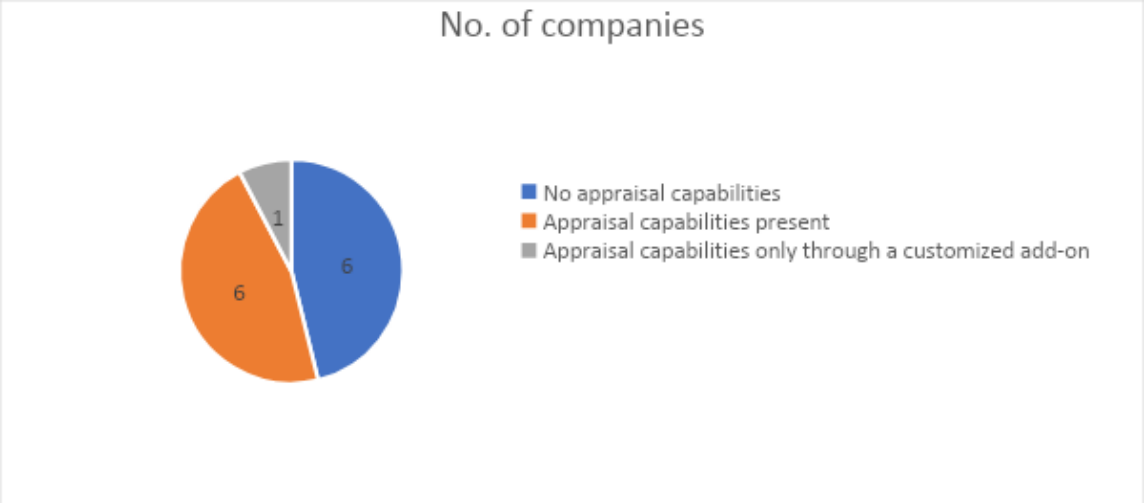

Figure 6. Appraisal capabilities

As to the second question about the identification of records to be disposed of by implementing a retention schedule (no. 12 of the questionnaire):

- Five companies replied that their products are not able to execute this activity (although one of them has added that "the document classifications and data - e.g., document date - that we provide are routinely used by RM systems for this purpose").
- One company replied that "It is possible using business rules. Required business rules can be written in the system to do so", which means that its platform in its original configuration (i.e., the "vanilla" platform) does not possess such a capability, but they can customize their product and enable it to perform this task.
- Seven companies answered that their platforms can identify records to be disposed of by applying a retention schedule, although two out of seven have not further elaborated on their reply.

As to the other five companies which said their products can identify records to be disposed of by applying a retention schedule:

- A company added that "we could extract dates and apply logic flagging documents if they were over a certain age", that means that the product does not directly apply retention schedules but can extract dates and – on the basis of the information it has acquired – can flag documents by comparing the dates with information given to it about particular deadlines.
- Another company clarified the ability to identify records to be disposed of and apply retention schedules is linked with the activity of classification and that "Appropriate disposal actions are assigned at the time of classification".
- A third company wrote its product can do so by "using workflows designed and implemented by users".
- Another responder said they have developed a "TRD file which regulates the times of the life-cycle of each document[14]".

[14] "TRD files are databases of file type definitions used by TrID, a software utility that can identify file types based on their binary signatures" cf. https://www.reviversoft.com/file-extensions/trd#:~:text=what%20is%20a%20..stored%20in%20the%20TRD%20format . Consulted on 31/05/2023.

- The last company of this group replied that its product can try to infer retention schedules by analysing the content, structure, and metadata of the documents.

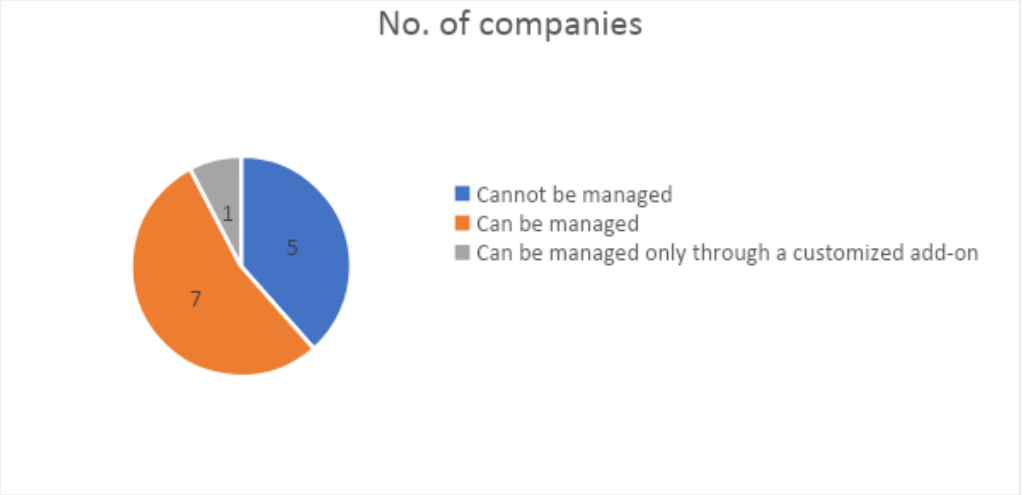

Figure 7. Retention schedules

To summarize the answers, the development of capabilities concerning appraisal and the implementation of retention schedules has not been so far a specific goal of the firms that have taken part in the survey, as the platforms of half of them do not possess at all such capabilities, while the other companies have linked such features to other characteristics of their products, such as classification, life-cycle management, workflow management and extraction of dates from the content of documents (although it is to be highlighted that linking appraisal and retention schedules implementation with classification and life-cycle management is an approach fully acceptable in archives and records management, at least for some schools of thought) or, in one case, to the deployment of additional modules to be added to the original platform.

Finally, it is to be noted that a company which replied its product can both try to appraise records and implement retention schedule has developed an application which is based on unsupervised learning and tries to build on its own accord its knowledge base, of course with the help of the input and feedback given by humans.

4.4 Technology solutions

4.4.1 Techniques and analysis models

A section of the survey was devoted to investigating the features of the technology enabling the AI-based products, and two questions of this section respectively dealt with the techniques and analysis models deployed by the companies to achieve the objectives their products have been designed for.

A question addressed the kind of analysis models used by the companies i.e., the kind of decision-making processes and general methodologies underpinning the work of their AI-based applications, while the second question investigated the specific techniques chosen by the companies.

With regard to the first question, the replies have been (between brackets the number of companies that gave a particular reply – of course several companies in their replies mentioned more types of analysis models):

- Neural Network models[15] (4 companies);
- Support Vector Machines a.k.a. State Vector Machines[16] (4 companies). One of the 4 companies have further specified they can use "*linear*" or "*probabilistic*" SVMs;
- Decision Trees[17] (3 companies);
- Random Forests[18] (3 companies);
- LSTM - Long short-term memory[19] (3 companies);
- Encoder-decoder[20] (2 companies);
- Convolutional Neural Network[21] (2 companies);

---

[15] "Neural networks—and more specifically, artificial neural networks (ANNs)—mimic the human brain through a set of algorithms. At a basic level, a neural network consists of four main components: inputs, weights, a bias or threshold, and an output." cf. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks Consulted on 29/05/2023.

[16] Support Vector Machines a.k.a. State Vector Machines "are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis" cf. http://algorithmtraining.com/state-vector-machines/. Consulted on 29/05/2023.

[17] "Decision trees in artificial intelligence are used to arrive at conclusions based on the data available from decisions made in the past. Further, these conclusions are assigned values, deployed to predict the course of action likely to be taken in the future." cf. https://www.upgrad.com/blog/decision-tree-in-ai/. Consulted on 29/05/2023.

[18] "Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models" cf. https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Consulted on 29/05/2023.

[19] "Long Short-Term Memory Networks is a deep learning, sequential neural network that allows information to persist. It is a special type of Recurrent Neural Network which is capable of handling the vanishing gradient problem faced by RNN. (...). Let's say while watching a video, you remember the previous scene, or while reading a book, you know what happened in the earlier chapter. RNNs work similarly; they remember the previous information and use it for processing the current input. The shortcoming of RNN is they cannot remember long-term dependencies due to vanishing gradient. LSTMs are explicitly designed to avoid long-term dependency problems." cf. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/. Consulted on 29/05/2023

[20] "In the field of AI / machine learning, the encoder-decoder architecture is a widely-used framework for developing neural networks that can perform natural language processing (NLP) tasks such as language translation, etc which requires sequence to sequence modeling. This architecture involves a two-stage process where the input data is first encoded into a fixed-length numerical representation, which is then decoded to produce an output that matches the desired format." cf. https://vitalflux.com/encoder-decoder-architecture-neural-network/, Consulted on 29/05/2023.

[21] "CNN is a type of deep learning model for processing data that has a grid pattern, such as images, which is inspired by the organization of animal visual cortex [13, 14] and designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. CNN is a mathematical construct that is typically composed of three types of layers (or building blocks): convolution, pooling, and fully connected layers. The first two, convolution and pooling layers, perform feature extraction, whereas the third, a fully connected layer, maps the extracted features into final output, such as classification. A convolution layer plays a key role in CNN, which is composed of a stack of mathematical operations, such as convolution, a specialized type of linear operation. In digital images, pixel values are stored in a two-dimensional (2D) grid, i.e., an array of numbers (Fig. 2), and a small grid of parameters called kernel, an optimizable feature extractor, is applied at each image position, which makes CNNs highly efficient for image processing, since a feature may occur anywhere in the image. As one layer feeds its output into the next layer, extracted features can hierarchically and progressively become more complex. The process of optimizing parameters such as kernels is called training, which is performed so as to minimize the difference between outputs and ground truth labels through an optimization algorithm called backpropagation and gradient descent, among others." cf.

- Transformer[22] (2 companies);
- Natural Language Processing (2 companies);
- Logistic Regression[23] (2 companies);
- Recurrent Neural Network[24] (1 company);
- TF / IDF - Term Frequency / Inverse Document Frequency[25] (1 company);
- Stochastic Dual Coordinated Ascent[26] (1 company);
- Single shot detectors[27] (1 company);

---

https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9#:~:text=CNN%20is%20a%20type%20of,%2D%20to%20high%2Dlevel%20patterns. Consulted on 29/05/2023.

[22] "A transformer model is a **neural network** that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence. Transformer models apply an evolving set of mathematical techniques, called attention or self-attention, to detect subtle ways even distant data elements in a series influence and depend on each other." cf. https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/. Consulted on 29/05/2023.

[23] "Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0" cf. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148. Consulted on 29/05/2023.

[24] "A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, natural language processing (NLP), speech recognition, and image captioning; they are incorporated into popular applications such as Siri, voice search, and Google Translate. Like feedforward and convolutional neural networks (CNNs), recurrent neural networks utilize training data to learn. They are distinguished by their "memory" as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depend on the prior elements within the sequence." https://www.ibm.com/topics/recurrent-neural-networks. Consulted on 29/05/2023.

[25] "TF stands for term frequency, or how often a term appears (that is, the density of that term in the document). The reason you care is because you assume that when an "important" term appears more frequently, the document is more relevant; TF helps you map terms in the user's query to the most relevant documents. IDF stands for inverse document frequency. This is almost the opposite thinking—terms that appear very frequently across all documents have less importance, so you want to reduce the importance weight of those terms." cf. https://www.infoworld.com/article/3339561/ai-machine-learning-and-deep-learning-everything-you-need-to-know.html?page=3#:~:text=TF%20stands%20for%20term%20frequency,that%20term%20in%20the%20document). Consulted on 29/05/2023.

[26] "In machine learning, the process of fitting a model to the data requires to solve an optimization problem. The difficulty resides in the fact that this optimization quickly becomes very complex when dealing with real problems. The Stochastic Gradient Descent (SGD) is a very popular algorithm to solve those problems because it has good convergence guaranties. Yet, the SGD does not have a good stopping criteria, and its solutions are often not accurate enough. The Stochastic Dual Coordinate Ascent (SDCA) tries to solve the optimization problem by solving its dual problem. Instead of optimizing the weights, we optimize a dual variable from which we can compute the weights and thus solve the former." cf. https://michaelkarpe.github.io/machine-learning-projects/sdca/#:~:text=The%20Stochastic%20Dual%20Coordinate%20Ascent,and%20thus%20solve%20the%20former. Consulted on 29/05/2023.

[27] "Single Shot Detectors (SSDs) are a popular and efficient method for object detection. They use a single convolutional neural network (CNN) to predict bounding boxes and class labels for objects in an image, making them faster and more efficient than other methods" cf. https://www.baeldung.com/cs/ssd#:~:text=In%20conclusion%2C%20Single%20Shot%20Detectors,more%20efficient%20than%20other%20methods. Consulted on 29/05/2023.

- Linear Regression[28] (1 company);
- K Nearest Neighbour[29] (1 company);
- Image processing (1 company, that have not further specified their reply);
- Deep Learning[30] (1 company, that have not further specified their reply);
- LSA - Latent Semantic Analysis[31] (1 company);
- Naïve Bayes models[32] (1 company);
- Conditional Random Fields[33] (1 company);

---

[28] "Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis." cf. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/#:~:text=Linear%20regression%20is%20an%20algorithm,machine%20learning%20for%20predictive%20analysis. Consulted on 29/05/2023.

[29] "k-nearest neighbor (k-NN) is one of the easiest and straightforward machine learning algorithms. It can be used for both regression and classification. It does not build a model unlike other machine learning algorithms; it does not have any trainable parameters. For every new test sample, it computes distances between this test sample and all training samples. Among all these distances, it chooses the "k" nearest training samples and then checks which class has maximum elements in the "k" closest set; it labels the test sample with the class having the maximum elements in the "k" closest set. The value of "k" is chosen empirically, it shouldn't be too large or too small. The selection of distance function is very important; it depends on the application." cf. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/k-nearest-neighbor#:~:text=Artificial%20intelligence%2Dbased%20skin%20cancer%20diagnosis&text=k%2Dnearest%20neighbor%20(k%2D,not%20have%20any%20trainable%20parameters. Consulted on 29/05/2023.

[30] "A neural network that consists of more than three layers—which would be inclusive of the inputs and the output—can be considered a deep learning algorithm." cf. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks. Consulted on 29/05/2023.

[31] "Latent Semantic Analysis, or LSA, is one of the basic foundation techniques in topic modeling. It is also used in text summarization, text classification and dimension reduction. (...) For LSA, we generate a matrix by using the words present in the paragraphs of the document in the corpus. The rows of the matrix will represent the unique words present in each paragraph, and columns represent each paragraph. The basic assumption for the LSA algorithm is that words that are closer in their meaning will occur in a similar excerpt of the text." cf. https://medium.com/acing-ai/what-is-latent-semantic-analysis-lsa-4d3e2d18417a. Consulted on 29/05/2023.

[32] "The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. (...) Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes. Naïve Bayes is also known as a probabilistic classifier since it is based on Bayes' Theorem.(...) This theorem, also known as Bayes' Rule, allows us to "invert" conditional probabilities (...) Bayes' Theorem is distinguished by its use of sequential events, where additional information later acquired impacts the initial probability. These probabilities are denoted as the prior probability and the posterior probability. The prior probability is the initial probability of an event before it is contextualized under a certain condition, or the marginal probability. The posterior probability is the probability of an event after observing a piece of data." cf. https://www.ibm.com/topics/naive-bayes#:~:text=The%20Na%C3%AFve%20Bayes%20classifier%20is,a%20given%20class%20or%20category. Consulted on 29/05/2023.

[33] "Conditional Random Fields or CRFs are a type of probabilistic graph model that take neighboring sample context into account for tasks like classification. Prediction is modeled as a graphical model, which implements dependencies between the predictions. Graph choice depends on the application, for example linear chain CRFs are popular in natural language processing, whereas in image-based tasks, the graph would connect to neighboring locations in an image to enforce that they have similar predictions." cf. https://paperswithcode.com/method/crf#:~:text=Conditional%20Random%20Fields%20or%20CRFs,implements%20dependencies%20between%20the%20predictions . Consulted on 29/05/2023.

- Passive Aggressive classifiers[34] (1 company);
- Association Rule Learning[35] (1 company);
- One of the companies replied that they have developed “New breakthrough techniques that perform the same tasks as the other AI models, but with the benefit of interpretability, fast computation and low energy footprint” and that they use “AI technique rather than a Machine Learning technique. This means that we are able to derive desired results without using ML techniques” (1 company).

It is really a wide range of models – the list above is made up of 24 entries, although in few cases the answers have been rather general without elaborating more on the particular models which have been adopted.

As to the second question – the types of techniques featured in the products of the companies, the answers have been (between brackets the number of companies that have given a particular reply – of course several companies in their replies have mentioned more types of techniques):

- Classification (9 companies);
- Clustering (5 companies);
- Extraction (2 companies; One of the companies further specified “feature and word extraction”);
- Topic Modelling[36] (2 companies; One of the two companies have clarified they use Topic Modelling for “keyword extraction”);
- Named Entity Recognition (2 companies);
- Regression[37] (2 companies);

---

[34] "The passive aggressive classifier is a machine learning algorithm that is used for classification tasks. (...) The passive aggressive classifier algorithm falls under the category of online learning algorithms, can handle large datasets, and updates its model based on each new instance it encounters. The passive aggressive algorithm is an online learning algorithm, which means that it can update its weights as new data comes in. The passive aggressive classifier has a parameter, namely, the regularization parameter, C that allows for a tradeoff between the size of the margin and the number of misclassifications. In each iteration, the passive aggressive classifier looks at a new instance, assesses whether it has been correctly classified or not, and then updates its weights accordingly. If the instance is correctly classified, there is no change in weight. However, if it is misclassified, the passive aggressive algorithm adjusts its weights in order to better classify future instances based on this misclassified instance." cf. https://vitalflux.com/passive-aggressive-classifier-concepts-examples/#:~:text=The%20passive%20aggressive%20classifier%20is,2006%20by%20Crammer%20et%20al. Consulted on 29/05/2023.

[35] "Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database." cf. https://www.javatpoint.com/association-rule-learning. Consulted on 29/05/2023.

[36] "Topic modeling is an unsupervised machine learning approach that can scan a series of documents, find word and phrase patterns within them, and automatically cluster word groupings and related expressions that best represent the set. Because it doesn't require a preexisting list of tags or training data that has been previously categorized by humans, this type of machine learning is known as 'unsupervised' machine learning. (...) Topic modeling is the method of extracting needed attributes from a bag of words. This is critical because each word in the corpus is treated as a feature in NLP. As a result, feature reduction allows us to focus on the relevant material rather than wasting time sifting through all of the data's text." cf. https://www.analyticssteps.com/blogs/what-topic-modelling-nlp. Consulted on 29/05/2023.

[37] "Regression finds correlations between dependent and independent variables. Therefore, regression algorithms help predict continuous variables such as house prices, market trends, weather patterns, oil and gas prices (a critical task these days!), etc. The Regression algorithm’s task is finding the mapping function so we can map the input variable of

- Generative Modelling[38] (1 company);
- Ranking (1 company);
- linguistic analysis (1 company);
- relationship analysis and subject-object relationship mapping (1 company);
- lexical analysis (1 company);
- inference algorithms (1 company);
- Handwritten Text Recognition (1 company);
- layout analysis (1 company);
- table recognition (1 company);
- One of the companies also mentioned – together with other techniques - "Natural Language Processing techniques based on fuzzy matching"[39] (1 company);
- One of the companies just replied that they use "AI technique rather than a Machine Learning technique. This means that we are able to derive desired results without using ML techniques" (1 company);
- One of the companies just answered, "It is a combination of many techniques to give better output" (1 company).

The list above is made up of 18 different entries, although – again – a few replies are rather general in their wording. Unsurprisingly – as the companies have been selected because they have a specific expertise at least in document management – classification (mentioned by 9 companies) and clustering (quoted by 5 companies) are the most reported techniques.

"x" to the continuous output variable of "y." cf. https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article. Consulted on 29/05/2023.

[38] "A generative model (...) concentrates on the distribution of a dataset in order to return a probability for a given occurrence.(...) In terms of a probabilistic model, a generative model specifies how a dataset is formed. We can produce new data by sampling from this model. Assume we have a dataset with images of cats. We might want to create a model that can create a fresh image of a cat that has never existed but still appears real because the model has learned the general rules that control a cat's appearance." cf. https://medium.com/codex/generative-models-the-next-machine-learning-boom-865b80c54fb1. Consulted on 29/05/2023.

[39] "Fuzzy matching (FM), also known as fuzzy logic, approximate string matching, fuzzy name matching, or fuzzy string matching is an artificial intelligence and machine learning technology that identifies similar, but not identical elements in data table sets. FM uses an algorithm to navigate between absolute rules to find duplicate strings, words/entries, that do not immediately share the same characteristics. Where typical search logic operates on a binary pattern, (i.e.: 0:1, yes/no, true/false, etc) – fuzzy string matching instead finds strings, entries, and/or text in datasets that fall in the in-between of these definitive parameters and navigates intermediate degrees of truth." cf. https://redis.com/blog/what-is-fuzzy-matching/. Consulted on 29/05/2023.

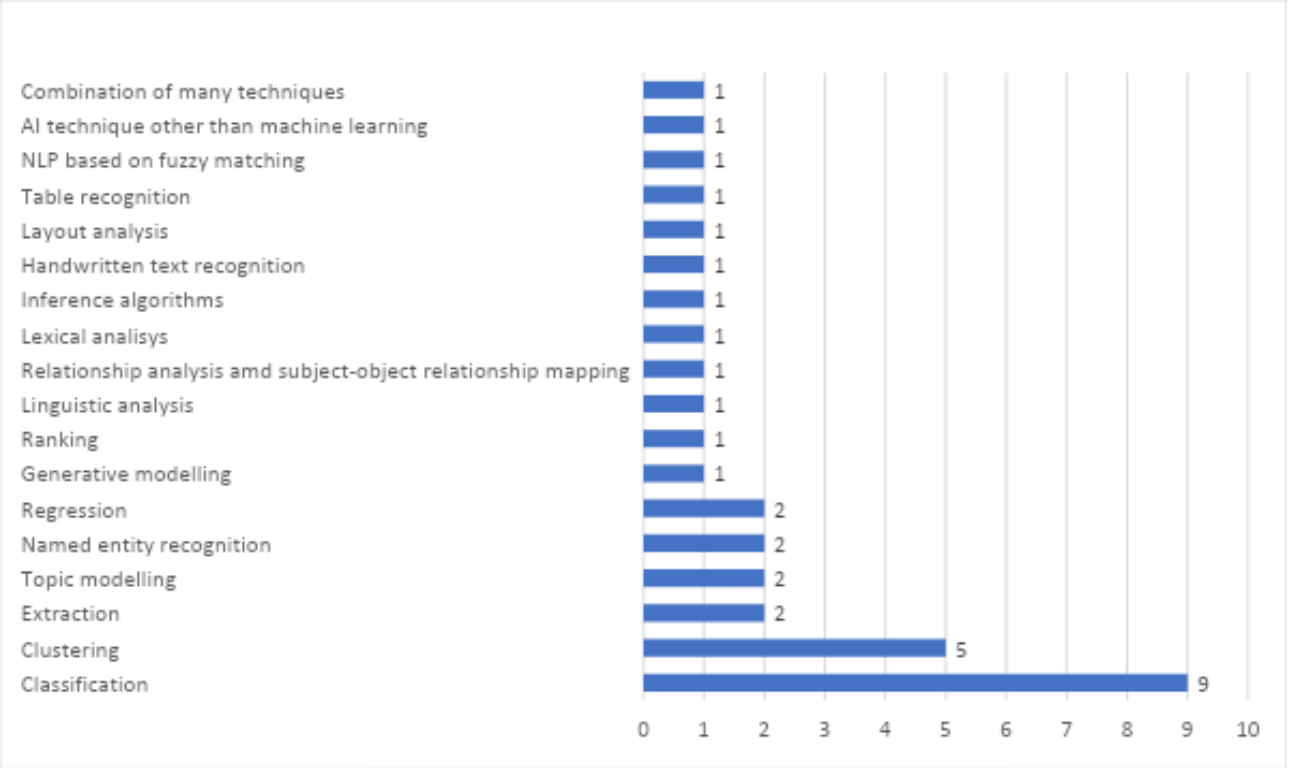

Figure 8. Techniques featured in the AI products

4.4.2 Training strategies

Team CU05 asked two questions in the interviews to understand how the companies make their AI-based applications "learn" to carry out the tasks they have been designed for.

One of the questions is just intended to categorise broadly the general strategies adopted by the companies: they were asked whether they use supervised, unsupervised, or semi-supervised learning[40].

The second question is more open-ended, as the companies were asked to describe the training process they set up (by specifying - e.g. - tools, methods, procedures) and to elaborate on how they select the sets of documents and data to train their products.

As to the first question, the outcomes are the following (of course, several companies in their replies have mentioned more kinds of training strategies among those listed below):

- Supervised Learning: 11 companies;
- Semi-Supervised Learning: 4 companies;
- Unsupervised Learning: 6 companies;

[40] "Supervised learning" means that humans train machines by feeding them labelled input and output data: in other words, humans show AI-based applications what they are expected to do.

"Unsupervised learning" is a strategy where an AI-based software works on its own to find out patterns, similarities and differences in data that have not been labelled by humans. Of course, also in unsupervised learning humans intervene to validate the outputs and "tell" the machine whether what it has found out makes sense or not.

In "Semi-supervised learning" AI-based products are fed at some point of the training process with a few labelled input and output data samples, and then they bring what they have learnt to bear on large sets of unlabelled data.

- Self-Supervised Learning[41]: 2 companies;
- Rule-based Learning[42]: 2 companies.

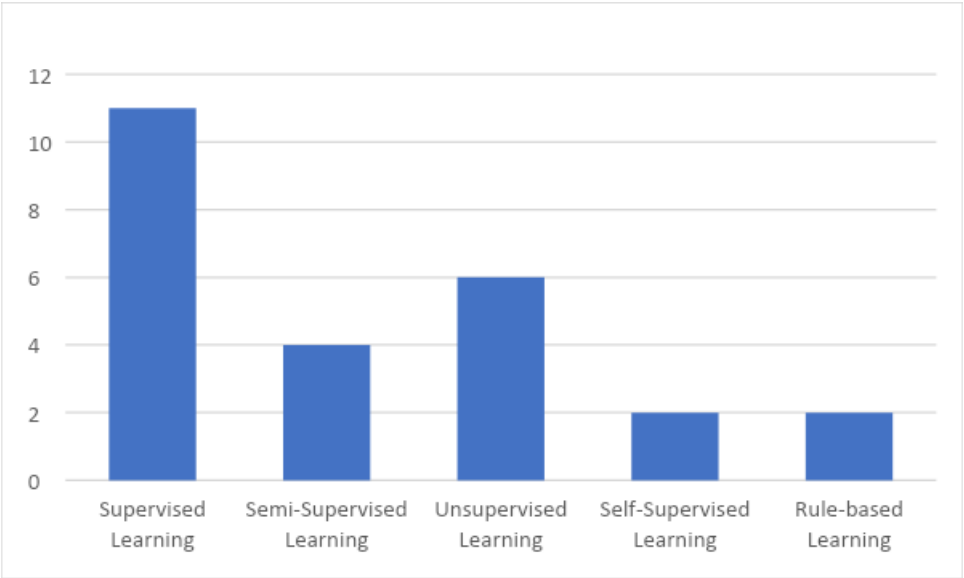

Figure 9. Kinds of training strategies

It is also interesting to see how many different training strategies the companies use. We found that:

- 6 companies use only one strategy, i.e.: 4 Supervised Learning; 1 Unsupervised Learning; 1 Rule-based Learning;
- 5 companies use two strategies, i.e.: 2 Supervised Learning and Unsupervised Learning; 2 Supervised Learning and Semi-Supervised Learning; 1 Supervised Learning and Self-Supervised Learning;
- 1 company uses three strategies, i.e. Supervised, Unsupervised and Semi-Supervised Learning;
- 1 company uses five strategies, i.e. Supervised, Unsupervised, Semi-Supervised, Self-Supervised and Rule-based Learning.

Please note that supervised learning is an approach adopted by almost all the companies CU05 team interviewed (i.e. 11 companies out of 13).

[41] "Self-Supervised Learning" (SSL) is a learning training model where an AI-based application processes an initial set on unlabelled data and generates on its own labels, that then are associated with new sets of unlabelled data; after an iteration, new labels are generated and then associated again with new sets of unlabelled data. In a sense it may be compared to "Semi-Supervised Learning", but in this case there is no human labelling of data at any stage of the learning process. SSL in some respects is also similar to unsupervised learning, but uses labels more than the latter does, although in this case the labels are created by the application itself and not by humans. Of course, also in SSL there is human interaction at a point in the learning process, as humans provide validation and quality improvement.

[42] Rule-based AI (a.k.a. Rules-as-Code a.k.a. Regulation-as-code) uses rules to solve a problem or carry out a task: since it works on specific rules coded by humans, the outcomes an AI application produces are pre-determined, as in this case AI models are based on conditional (i.e., 'if-then') statements. It goes without saying that in rule-based AI the role of human expertise and intervention is far more crucial and compelling than it is in machine learning.

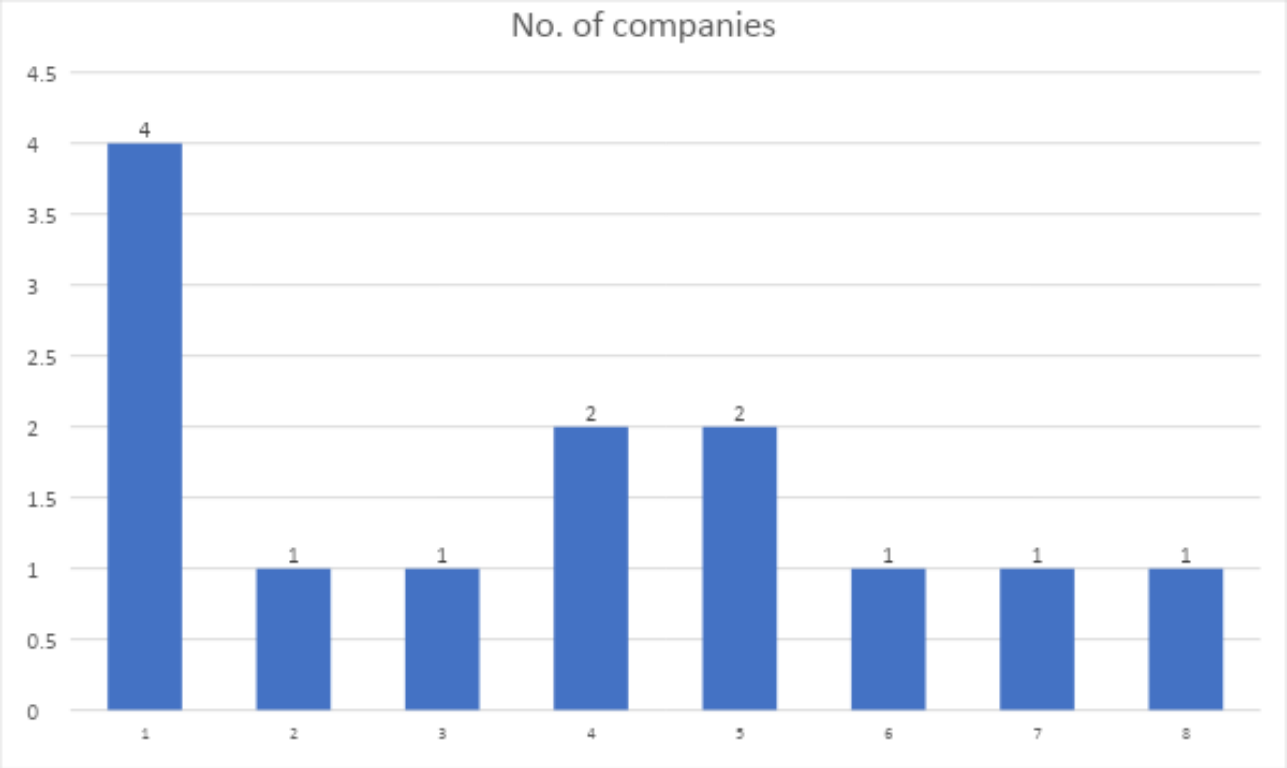

Figure 10. How many and which strategies are used

As to the second question, the answers that were given are understandably rather different from one another, since in many of the companies a specific procedure has been set up to carry out the training:

- 1 company declared they do not train their application on existing data, as their rule-based approach basically means that it is humans who directly "explain" to the AI-based product how to process the data;
- 2 companies use pre-trained models[43]. One of the two companies then also use "few samples of documents covering most of the variations" to refine the training;
- 2 companies said they train the product by using data and documents provided by their customers; one of these two companies, however, added that the raw data provided by the customers is enriched by the staff of the company "with metadata (i.e., grammatical and semantic information) coming from the core components of the technology: a disambiguation engine and a general-purpose Knowledge Graph (i.e., a lexical database where concepts are arranged based on semantic relations like loose synonymy)";
- 1 company answered that it is up to users to train the application: "Mainly users are creating the training data themselves and train also their AI models themselves";
- 1 company replied it is up to users to provide training samples, but they make available various tools "to speed up the training and testing process" such as "clustering, dynamic online learning (types will be suggested for all documents after only a few have been trained), outlier detection, generation of confusion matrices and classification/accuracy curves";
- 1 company initially builds a knowledge database by asking their customers relevant materials (according to what the AI application is supposed to do) that can be complemented by other

[43] Pre-trained models are models that have previously been trained on a large dataset: they can be saved and be subsequently used by developers of AI applications.

documents added by the staff of the company. Then two different options may be implemented to carry out the training: 1) either each document of the knowledge base receives an evaluation score and then the AI product is trained on all the documents, or 2) the documents of the knowledge base are randomly split into two groups, one to train the application and the second to be used as evaluation data (in this case, usually 80% of the documents are used as training data and 20% of documents are used as evaluation data);

- 1 company wrote they use the VGG16 deep network[44], which they train by means of "a *large volume of images*";
- 1 company answered they train their application by selecting documents to be used as positive or negative examples of a given class;
- 1 company creates a large database by using texts extracted from various containers, such as word, pdf and txt files. The information concerning the relationship of the text to the container is retained. Afterwards there is a stage of verification of the data and then the training of the application on the database starts off by using a Self-Supervised Learning approach;
- 1 company wrote that they use "an algorithm that selects the training set to favour recent records with a minimum and maximum amount of text data available. The training set is automatically supplemented with misclassified records so it can learn from its mistakes. Models based on the training set are selected using K-Fold cross validation[45]";
- 1 company basically replied that they adapt the training process to the specific case they have to deal with, as they said that "depending on the requirements, we approach the problem with right problem definition, assumption, data collection and labelling, model selection, training, evaluation with the right accuracy metrics".

Five companies (out of 13) said that their customers play a role in the training of the application, although the role may be more or less significant: that is unsurprising, as the training of an AI-based application will be more effective if someone conversant with the domain-specific knowledge necessary to understand the documents analysed by the application takes part in the training process.

It is to be noted that machines also "learn" what they should not do: one company feeds the AI-based application negative examples, another one uses "misclassified records" to let the machine "learn from its mistakes", a third company assigns each document used to train the machine an evaluation score, which - of course - may also be a low evaluation score.

4.4.3 Information elements processed by the platforms

A third set of questions addressed the information elements the AI-based products of the surveyed companies use to analyse and process the documents. More in particular, Team CU05 asked:

- whether any specific elements in the structure, form and content of a document are considered by the applications and - if this is the case - which ones;
- whether any specific metadata elements of a document are considered by the applications and - if this is the case - which ones.

---

[44] VGG is a Convolutional Neural Network (CNN) model supporting 16 layers - VGG stands for Visual Geometry Group. A Convolutional Neural Network is a kind of neural network to process data that has a grid-like topology, such as an image.

[45] "Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation". Cf. https://machinelearningmastery.com/k-fold-cross-validation/. Consulted on 27/05/2023.

CU05 Team also made it clear that - by using the words “metadata elements of a document” - we meant elements not directly present in the structure, form, and content of the document itself.

Regarding the information elements that can directly found in a document:

- 1 company replied that - with regard to what its application can use - “everything can be considered: text, layout, tables, etc...”;
- 4 companies said their products are able to analyse any part of the content of a document;
- 1 company answered its platform examines particular kinds of content, such as - e.g. - key phrases and named entities;
- 5 companies declared their products analyse the structure and form of a document and more in particular:
    - 1 company wrote its platform “parses the document structure and can use it as part of the model training. Some of the structure features used comprise pages, titles, headers, font size and words capitalization”;
    - 1 company stated they use document structures for categorising documents (such as - e.g. - CV, Bills);
    - 1 company just remarked that its product “uses the structure of documents and context of data to generate output”;
    - 1 company answered that – when needed – its application can parse extrinsic elements of the documentary form such as - e.g. - page size, colour, layout;
    - The fifth company replied that its products can examine the layout of documents even if that “mainly concerns formatting properties such as the font (bold or italic), titles and tables”;
- 2 companies said that what is analysed depends on the circumstances: one company just wrote that they can configure what their product takes into consideration, while the other company said that what their application examines “depends on the customer requirements”, without adding more details.

As to the metadata elements used by the applications:

- 4 companies declared that they can use any kind of metadata elements found in a document; more in particular: 1 company said “this also includes metadata added by AI enrichment, external systems, or signal calculations”; 1 company clarified the metadata elements also encompass “location, document type, date-time modifiers (last modified, created date), and any custom metadata fields within SharePoint Online”; the other 2 companies did not add any additional information.
- 1 company wrote that “the metadata items can be chosen by the user and are project-specific”;
- 1 company replied that “the metadata collected varies by the document type” and - e.g. - for emails their platform collects “file name, tagged date, creation time, subject, from, to, source”; for MS Word documents “creation time, source, file name, file path, tagged date”; for PDF documents “author, title, subject, creation time, total page number”; for MS PowerPoint files “author, category, comments, content status, creation time, identifier, keywords, language, modified, subjects, title, version”;
- 1 company answered its application collects “author (origin), date and recipient”;
- 2 companies declared that their products can use metadata without elaborating on their replies: one of the two companies just added their application can analyse the “*relationships*” of a document;

- 1 company only replied that what its application examines “depends on the customer requirements”, without adding more details;
- 2 companies stated their platforms focus on the content of documents and as a rule do not consider metadata elements, but also added that if needed it is possible to configure their products to collect and parse metadata elements, without giving more details;
- 1 company just answered its platform does not analyse metadata.

As you can see, the landscape emerged from the interviews is somewhat assorted, as the thirteen companies that were surveyed look at various combinations of information elements. In several cases the companies did not specifically list the components of a document or the metadata elements they analyse, but made general statements, which on the one hand is understandable as flexibility and ability to cater for the particular requests of a customer are clearly assets for these players, on the other hand may also be a sign there has been not time yet to develop and establish standard business processes in this field: the state of affairs is shifting and there is room to shape future steps and set priorities for those who have the strength to exert influence on the agenda of this industry.

4.4.4 Affordances and constraints of the IT ecosystems

Team CU05 decided to ask the companies a question about both the constraints their respective AI-based products require the users to comply with and the resources the same products enable the customers to work with i.e., the IT environments and systems their applications can interact with. In other words, we asked the companies to describe the IT ecosystems their products need to operate properly and create value.

1 company did not give any reply to this question, and as to the other 12 companies (NB: please note that a company in their reply may have mentioned more solutions and/or issues among those listed below):

- 5 companies said there are some constraints for their products:
    - § 1 company wrote that its platform “requires MS windows and SQL Server as a backend”;
    - § 1 company specified the list of document management systems its AI-based products can connect with – which implies that at the moment their products need one of these systems to run properly. They also added that at the moment there are “some limitations with thresholds and volume”, and that they “are aiming to expand the number and type of content sources” their products can connect to;
    - § 1 company observed that - although by developing APIs they can enable their product to interact with a wide range of platforms – “some content sources do not allow external programs to delete (dispose) data”, and this of course is a constraint on the actions their product can perform;
    - § 1 company specified its product is “capable of direct interaction with almost all web-enabled systems” (but not with any kind of web-enabled system).
    - § 1 company said its application is only available on cloud.
- 2 companies simply said that there are no technological barriers for their products without adding any further detail;
- 1 company wrote they “have developed Web Services to get integration with any other applications” and therefore there are no technological barriers for their AI-based application;
- 4 companies replied that they use APIs (i.e., Application Programming Interfaces) to enable their respective products to interact with various platforms and IT environments. One of the companies added that this approach allows them to enable “direct interaction with almost all web-enabled systems”. Another company also specified that “the API can also be independently deployed where

it is needed to support functionality from another application". A third company stated that their AI-based application "can be integrated with a wide variety of frameworks, because it exposes a REST API[46] with all the product capabilities";

- 3 companies answered that they can let their products interact with external solutions through customized connections they are able to build: one of the three companies said they "can interact with more than 300 external solutions because we have created a 'connector factory' in our platform", the second company wrote that their "technology can be deployed across most modern IT infrastructures, on premises or on the cloud, in Windows-based or Linux-based environments, with no specific limitation, thanks to the availability and ease of customization options to connect to existing systems via standard and proprietary formats and protocols", the third company replied that they can "manage data from On Premises systems such as file shares and on premises SharePoint" by "usually working with customers' IT administration to ensure that their security protocols are being met";
- 1 company stated that its platform – which uses Natural Language Processing – can access "1600 different file formats to extract key phrases and named entities" and that so far they "have not encountered records of business we could not parse" yet;
- 1 company wrote that they are also able to support "looser, file-based integrations" and that "for non-cloud usage, direct code-level integration to the technology is possible via a .NET SDK". They also clarified that "there are no requirements for specific IT systems or 3rd party software" with regard to the possibility of developing integrations.

To sum up, most of the AI-based products that have been reviewed can thrive in every operating system (Windows, Linux, Mac, etc.) and interact with – as a minimum – with a very large number of platforms and applications by developing customized APIs and connectors, which of course takes an effort and requires resources. Moreover, almost all the AI-based products reviewed can be deployed and work both on premises and on cloud.

5 PERFORMANCE MEASUREMENTS

An essential stage in every business process is to find a way to evaluate whether and to what extent the objectives that had been set out have actually been achieved. To that end, it is necessary to find parameters to quantify the performance levels of a product. Team CU05 therefore decided to ask the companies which kind of metrics they have chosen to measure the success rates of their applications.

Moreover, our Team also asked each company whether they had devised any solution to tackle the problem of algorithm biases that affects various AI-based software programs – Archives and records management may easily have a deep impact on human lives and it is important to make sure AI-based products may not undermine rights, foster inequalities or simply wreak havoc: hence the ability to detect of possible algorithm biases is an additional element for performance measurement.

As to the first question (i.e., kinds of adopted metrics), these are the metrics that have been found in the replies (between brackets the indication of how many companies have adopted a particular kind of metrics – of course, several companies in their replies have mentioned more kinds of metrics among those listed below):

[46] A REST API is an API that conforms to the design principles of the REST, or REpresentational State Transfer architectural style.

- F1 score. F1 score is a machine learning evaluation metric that measures the accuracy of a model: it combines the precision and recall scores of a model. The accuracy metric calculates how many times a model made a correct prediction across the entire dataset (5 companies). One of the companies said that the expectation is usually for 80% accuracy or 0.80 F1 score;
- The precision-recall curve which shows the trade-off between precision and recalls for different thresholds. The "precision" is referred to the proportion of correct predictions among all predictions for a particular class and the "recall" is referred to the proportion of examples of a particular class that has been predicted by the model as belonging to that class - therefore high precision relates to a low false positive rate, and high recall relates to a low false negative rate, i.e. the higher the scores, the better the performance (4 companies);
- the Area Under the Curve - Receiver Operating Characteristic curve (AUC-ROC). ROC is created by charting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, and AUC is a method to evaluate the accuracy of the outcomes: the closer the ROC curve is to the upper left corner of the graph, the higher the accuracy of the outcomes because in the upper left corner of the graph the sensitivity is 1 and the false positive rate is 0. The Receiver Operating Characteristic curve is rather similar to the precision-recall curve (2 companies have said they have adopted AUC-ROC curve);
- Mean Average Precision (2 companies);
- the Matthews correlation coefficient (MCC), which produces a high score only if the prediction obtained good results in all of the four categories of a matrix - i.e., true positives, false negatives, true negatives, and false positives - in proportion to both the size of positive elements and the size of negative elements in the dataset (1 company);
- Sampling, by carrying out a human-led, blind sentencing activity and then by comparing the machine results (1 company). The company that have adopted that approach have set the accuracy benchmarks to achieve to 98-99%;
- Character Error Rate (1 company);
- Word Error Rate (1 company);
- "How many correct answers we give to users divided by total questions" (1 company - approach chosen by a company using Question-Answering Models to assess the performance. Question-Answering Models are machine or deep learning models that can answer questions given some context or also without any context);
- One company answered that "Results are compared to known values for datasets for which a previous annotation of expected results is performed (these datasets are known as 'golden standards' or 'ground truth'). The evaluation processes are integrated in the platform to allow for continuous monitoring of quality to ensure full control on the system performance" (1 company);
- The number of documents that have been processed by the AI-based platform (1 company);
- Ingestion thoroughness (1 company);
- Ingestion speed (1 company);
- Classification coverage (1 company);
- Acceptance rates for Machine Learning-based classifications (1 company);
- Number of overdue disposal actions (1 company).

It is worth noting that 3 companies also replied that they use application-specific metrics defined by their customers or agreed on together with their customers, and that 4 companies added they use several other kinds of metrics (without listing all of them in their replies). Finally, 1 company did not give any reply to this question.

As to the question concerning the action to forestall and detect algorithm biases, the following replies were given (between brackets the indication of how many companies have adopted a particular kind of metrics –

of course, several companies in their replies have mentioned more kinds of metrics among those listed below):

- Confidence, range, and probability scores assigned to classifications or other kinds of operations (2 companies). One of the companies that gave this reply also added that this solution is currently being built up by their Research and Development Department. The other company that answered the same way added that if the assigned scores are too low humans intervene and carry out the classification or operation;
- Bayesan inference, that is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. This methodology may allow users to control bias (1 company). The company that have adopted this approach wrote in their reply that "For instance, we can emphasize longer documents that contain a rich usage of financial vocabulary. This avoids erroneous classifications, for instance, like classifying a document as financial just because it contains a currency value";
- Identification of misclassified records that are then fed back into the machine learning model to help overcome biases (1 company);
- One of the companies answered that "Systems are placed into use after training on the user-specific data, hence chances of algorithm biases are very less" (1 company);
- Sentiment analysis. Sentiment Analysis (a.k.a. opinion mining) is a Natural Language Processing technique used to determine whether data is positive, negative or neutral and is performed on textual data to help users monitor brand and product sentiment in customer feedback, and understand customer needs (1 company);
- Tools based on TF-IDF (Term Frequency - Inverse Document Frequency), which is an algorithm that uses the frequency of words to determine how relevant those words are to a given document (1 company);
- Identification and targeting of documents which have specific characteristics and may be more subject to generating biases - e.g., documents that have lots of words but little semantic content (1 company);
- Deep long-tailed learning, which aims to train well-performing deep models from a large number of images that follow a long-tailed class distribution[47] (1 company). The company that have given this reply have also added that this solution is currently being built up by their Research and Development Department;
- Direct checks to find biases (1 company).

To be noted that 4 companies said they do not undertake any specific action to tackle algorithm biases, 2 companies replied that they are very confident the technologies used to develop their products will be able to prevent algorithm biases and 2 companies wrote they rely on their metrics to measure performance to detect possible algorithm biases. Finally, 2 companies added they are currently still working to find solutions for algorithm biases.

By and large, the general feeling by reading the replies to the question about algorithm biases is that this issue is very complex for the surveyed companies and that there are not well-established procedures and tools yet to deal with this problem.

[47] "In a long-tailed distribution, a small proportion of classes account for the majority of data, while most of the other classes lack enough data to be representative". Cf. H.Zhao, S.Guo, Y.Lin "Hierarchical classification of data with long-tailed distributions via global and local granulation" in *Information Sciences*, Volume 581, December 2021: 536-552, https://www.sciencedirect.com/science/article/abs/pii/S0020025521009968. Consulted on 05/06/2023.

6 FINDINGS

6.1 Remarks from the archival perspective

The effort made in identifying and selecting market solutions provided a list of companies able to understand the complexity and the relevance of archival functions aimed at supporting the records relations in connection with the business context. The replies to the questionnaire generally testify a common awareness of the central role of the specific original metadata created in the creator's current activities, both if the issue concerns the records' automatic classification or in case of the creation of archival aggregations. Of course, this perception is even stronger when the action implies the re-reconstitution of archival original contexts. For this reason, the presence of any metadata fields found or inferred on records is at the centre of any replies. The records typology – when available – is often considered another crucial component for the successful application of the AI techniques to the records. In terms of records classification, only one company pointed out the capacity of its platform to be trained by the users, thanks to a specific set of data for generating autonomously labels and tags related to any record classification scheme understood as based on taxonomy or term ontology. In the other cases the human intermediation seems still unavoidable for providing consistent results.

In terms of records aggregation or re-aggregation, the promises for automatization are not very encouraging, as this possibility is limited to very specific cases: for document types, when the users' specifications are in place, or the structure of the content source provides basic information. The automatic or semi-automatic aggregation based on the document content is only suggested with the support of the user validation, of human-in-the-loop workflow or when content rules are created and enabled. In more cases even these capacities are not already developed but in the process of being developed.

Even the provenance information seems not easily recognisable by AI solutions when based on inferences and without very specific requirements (such as the identification of the right case-folder, the presence of a stamp, a statement clearly expressed in the record, specific metadata and/or classification elements).

Consistently with previous analysis, also the reconstitution of the archival bond – when lost or not explicitly defined – is not a simple and easy activity to be dealt with by AI solutions, without the significant help of users and/or consistent descriptive information available and, in any case, it implies more investments, not yet supported by the market.

A similar observation can be made for appraisal, not really developed by the companies which accepted the interview. The only positive answer related to a product usable for appraising records or implementing retention schedules admits that the process requires the help of the input and feedback given by humans.

In conclusion, we have noticed a general cautious approach in all the replies when the questions were related to the records and archival contextual relations. Of course, these remarks imply further analysis, as many other market proposals for archival and records management are not in line with this evaluation. The reasons for this gap could depend on the strict parameters we have adopted for selecting the market solutions, but also it could relate to the degree of interactions and explanations exchanged between the researchers and the companies involved in the review during the questionnaire submission. In any case, it testifies that the complexity of our functions, at the moment cannot be easily reduced and removed by an automatic approach, but only supported by the AI technologies through the intermediation of users and professionals. We are not able to say, without further analysis and case studies, which degree of professional intermediation is and will be necessary at least in the next and medium-term future. More effort is still required for assessing and measuring the quality and the consistency of new AI tools and their promises for automatically supporting or even substituting the human activities for classifying and

aggregating records on a functional, accurate and reliable basis. This is an essential reason for our study group to address our efforts on case studies with the aim of acquiring concrete elements for understanding how the archivists could provide their support in this crucial phase of digital transformations.

6.2 Remarks from the technical perspective

The survey contained some questions about the technical solutions adopted by the companies, and more specifically about analysis models and types of techniques used in the products; training strategies; information elements processed by the platforms; and the features of the IT ecosystems the AI-based products need to operate properly.

As to the analysis models, the landscape is really varied: companies declare to use overall 24 different models. Neural Networks (7 times) and Support Vector Machines (4 times) are the most mentioned models. It should be noted that in some cases the answers have been rather general.

Regarding the types of techniques used by the companies' products, Classification (9 times) and Clustering (5 times) are the most common answers. This was to be expected since the companies were selected for their expertise in document management. 18 different types of techniques were mentioned overall, and in this case a few replies were general without any elaboration.

As for training strategies for AI-based products, the survey showed a mixed situation: 11 companies use supervised learning, 6 companies use unsupervised learning, 4 companies use semi-supervised learning, 2 companies use self-supervised learning, and 2 companies use rule-based learning. Of course, several companies use more than one type of training strategies.

The answers to the questions about the information elements processed by the AI-based applications again showed a multifaceted situation: the thirteen companies that have been surveyed look at various combinations of information elements (e.g. text, layout, tables, page size, colours, various kinds of metadata elements).

It is worth noting, however, that in several cases the companies did not specifically list the components of a document or the metadata elements they analyse, but just made general statements.

All that is likely to mean that the state of affairs is shifting and there is room to shape future steps and set priorities for those who have the strength to exert influence on the agenda of the industry of the AI-based applications.

As to the characteristics of the IT ecosystems required to run the AI-based products, in a nutshell we can say that most of the AI-based products that were reviewed can thrive in every operating system (Windows, Linux, Mac, etc.) and interact with – as a minimum – a very large number of platforms and applications thanks to customized APIs and connectors, whose development takes an effort and requires resources.

Moreover, almost all the AI-based products reviewed can be deployed and work both on premises and on cloud.

ANNEX 1 – THE QUESTIONNAIRE

INTERVIEW TO COMPANIES USING AI FOR THE ARCHIVAL DOMAIN

NAME OF THE COMPANY: DATE:

I SECTION: ACHIEVEMENTS

1. Please list and describe application(s) you have developed for archives and records management
2. What type of platforms are the application(s) developed for (e.g., Business information systems; Filing systems; ERMS; EDMS; Email client applications; Intranet; Web archiving; Social media; Messaging applications; Video conferencing applications; Databases)?
3. Please describe what has been achieved through the use of your application(s) (i.e., the portfolio of your application(s))
4. Can you describe the main features and strengths of your applications and - if any – future plans to develop it?
5. Are there aspects you want to improve and/or problems to be solved?
6. Have you ever cooperated with archival institutions and/or university departments / research centres involved with archives and records management?
7. Which archival and records management standard are you in compliance with, if any (e.g., ISO 15489, DoD, Moreq etc.)?

II SECTION: SPECIFIC CAPABILITIES (FOR RECORDKEEPING AND EMAIL SYSTEMS)

8. Is your application able to file automatically or semi-automatically records in the respective folders, case-files or group they may belong? May it perform this task for both newly created records and accumulated records?
9. Is your application able to classify/file records, folders and groups of records based on a records classification scheme in which functions, administrative processes, document type are identified?
10. Is your application able to extract metadata from records and use these metadata to describe them, even when records contain hand-written text? In case of a positive answer, please provide examples
11. Is your application able to carry out appraisal and identify records with archival value?
12. Is your application able to identify records to be disposed of, based on a records retention schedule?
13. Is your application able to re-constitute archival aggregations that have been lost? In case of a positive answer, can you clarify the process and provide examples?
14. Is your application able to index records in order to provide information about related links or aggregations among records?

III SECTION: TECHNOLOGIES AND METHODS USED IN THE AI APPLICATIONS

15. Which types of models do you use to support Artificial Intelligence (e.g., neural network models; gaussian mixture models; latent Dirichlet allocation; encoder-decoder; long short-term memory; etc.)?
16. Which kinds of strategy do you use: supervised, semi-supervised or unsupervised?
17. Which kinds of specific techniques do your applications use (e.g., clustering; classification; regression; topic modelling; generative modelling; etc.)?
18. Which kind of training strategy have you chosen for your applications (please specify tools, methods, procedures etc.) and how do you select the sets of documents and data to train your application?
19. Which, if any, elements in the structure, form and content of a document are considered by your application(s) to make decisions?
20. Which, if any, metadata elements of a document are considered by your application(s) to make decisions? (By using the term metadata elements, we mean elements not directly present in the structure, form and content of the document itself);
21. Is your application able to make inferences about which records belong or might belong to the same group or business process (e.g., same case-file, subject-file, series, fonds)?
22. Is your application able to make inferences about the organization or person that has created or received and then set aside the records, even when relevant metadata elements for their identification are missing?
23. Which IT environments and systems is your application able to interact with? Are there technological barriers to its actions (e.g., specific software or proprietary formats)?

IV SECTION: AUDIT- CHECKS - KEY PERFORMANCE INDICATORS

24. Which kinds of metrics do you use to measure the success rates to achieve the objectives your applications have been designed for (e.g., automatic classification / indexation, intelligent discovery, automatic redaction, automatic implementation of retention schedules or whatever else your applications are meant to do)?
25. Have you devised any solution to identify algorithm biases that can impact on the outcomes of your application (e.g., short documents containing calculations involving currency values might automatically be classified as financial documents, while might be estimates in a legal action or short report of an important project)?