

Archivi e intelligenza artificiale responsabile

Esperienze,
opportunità, cautele

16 maggio 2025



10.00 Aula Magna
13.00 Regione Emilia-Romagna, viale Aldo Moro 30

> 9.30 Registrazioni e caffè di accoglienza

> 10.00 Saluti istituzionali

Elena Mazzoni *Assessora Regione Emilia-Romagna*

Francesco Raphael Frieri *DG Regione Emilia-Romagna*

Gabriele Bezzi *Presidente ANAI sezione Emilia-Romagna*

Matteo Sisti *Presidente ANAI sezione Marche*

Fabrizio Monti *Presidente EBLA soc. coop.*

Introduce e presiede **Allegra Paci** *EBLA soc. coop.*

> 10.30 **Giorgia Di Marcantonio** *Università di Napoli, l'Orientale*
INTELLIGENZA ARTIFICIALE E ARCHIVI: DI COSA PARLIAMO DAVVERO

> 11.00 **Paolo Torroni** *Università degli studi di Bologna*
L'USO DELL'AI NELLA CLASSIFICAZIONE AUTOMATICA DI TESTI

> 11.30 **Marianna Tascone** *Regione Emilia-Romagna*
CALCOLO AUTOMATICO DELL'INDICE DI INTEROPERABILITÀ DEI FORMATI

> 12.00 **Stefano Allegrezza** *Università degli studi di Macerata*
CLASSIFICARE E FASCICOLARE I DOCUMENTI CON L'INTELLIGENZA ARTIFICIALE: LUCI ED OMBRE

> 12.30 **Pierluigi Feliciati** *Università degli studi di Macerata e ANAI Marche*
IL PROGETTO INTERNAZIONALE INTERPARES TRUST AI (2021-2026) E IL SYMPOSIUM A BOLOGNA DEL 20 GIUGNO 2025

Light lunch

15.00 Cappella Farnese Palazzo d'Accursio
17.30 piazza Maggiore 6, Bologna

> 15.00 Saluti istituzionali

Roberta Li Calzi *Assessora Comune di Bologna*

Elena Stefani *Soprintendente archivistica e bibliografica dell'Emilia-Romagna*

Introduce e presiede **Gabriele Bezzi** *ANAI sezione Emilia-Romagna*

> 15.30 **Carlo Batini** *Università degli Studi di Milano-Bicocca*
VERSO UN LINGUAGGIO DI COMANDI COGNITIVI BASATO SU MODELLI LINGUISTICI PER ELABORARE I DOCUMENTI DEL FASCICOLO DEL PROCESSO CIVILE

> 16.00 **Francesca Delneri** *Archivio di Stato di Bologna*
NON DI SOLA DIGITALIZZAZIONE. ESPERIENZE DI DESCRIZIONE, RIPRODUZIONE E ACCESSO AI FASCICOLI DI CORTE D'ASSISE CONSERVATI DALL'ARCHIVIO DI STATO DI BOLOGNA

> 16.20 **Riccardo Righi** *Regione Emilia-Romagna*
COME MIGLIORARE LA FRUIZIONE DEI CONTENUTI D'ARCHIVIO ATTRAVERSO L'INTELLIGENZA ARTIFICIALE: UNA SPERIMENTAZIONE SUI FASCICOLI PROCESSUALI RELATIVI ALLA BANDA DELLA UNO BIANCA

> 16.45 Discussione e conclusioni

Conclude **Claudio Leombroni** *coordinamento MAB*

con il patrocinio di



**Comune
di Bologna**

Iscriviti con il QR code

o [clicca qui](#)

utente: iscrizione

passw: iscrizione



Per maggiori informazioni: segreteria@eblacoop.it



InterPARES Trust^{RI}



unIMC
UNIVERSITÀ DI MACERATA

DIPARTIMENTO DI
STUDI UMANISTICI
Lingue, mediazione,
storia, lettere e filosofia

Classificare e fascicolare i documenti con l'intelligenza artificiale: luci ed ombre

Stefano Allegrezza
Università degli Studi di Macerata
stefano.allegrezza@unimc.it



Agenda

1. Introduzione
2. La ricerca condotta sull'impiego dell'AI per la classificazione e fascicolazione
3. Risultati
4. Considerazioni finali

1. Introduzione

La classificazione dei documenti informatici

[Linee guida AgID, par. 3. Classificazione dei documenti informatici]

La **classificazione** ha il fine di organizzare logicamente tutti i documenti amministrativi informatici prodotti o ricevuti da un ente nell'esercizio delle sue funzioni. L'attività di classificazione si avvale del **piano di classificazione** che mappa, su più livelli gerarchici, tutte le funzioni dell'ente.

La **classificazione** è un'**attività obbligatoria** nel sistema di gestione informatica dei documenti dell'AOO e si applica a tutti i documenti prodotti e acquisiti dalla stessa AOO sottoposti o meno alla registrazione di protocollo, ai sensi degli articoli 56 e 64, comma 4, del TUDA. Le informazioni relative alla classificazione nei casi dei documenti amministrativi informatici costituiscono parte integrante dei metadati previsti per la formazione dei documenti medesimi.

Nel sistema di gestione informatica dei documenti dell'AOO l'**attività di classificazione** guida la formazione dell'archivio mediante il **piano di organizzazione delle aggregazioni documentali**.

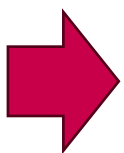
[Linee guida AgID, All. 1 «Glossario dei termini e degli acronimi»]

Piano di classificazione: «Struttura logica che permette di organizzare documenti e oggetti digitali secondo uno schema desunto dalle funzioni e dalle attività dell'amministrazione interessata».

Gruppo di lavoro per la formulazione di proposte e modelli
per la riorganizzazione dell'archivio dei Comuni

Piano di classificazione (= Titolario)
per gli archivi dei Comuni italiani
(seconda edizione)

dicembre 2005



Senza la classificazione non si produce un archivio ma un insieme di documenti senza alcuna relazione tra loro, che rende difficile lavorare e inefficiente la ricerca

La formazione delle aggregazioni documentali

[Linee guida AgID, par. 3.3 «Aggregazioni documentali informatiche»]

La Pubblica Amministrazione documenta la propria attività tramite funzioni del sistema di gestione informatica dei documenti finalizzate alla produzione, alla gestione e all'uso delle **aggregazioni documentali informatiche**, corredate da opportuni metadati, così come definiti nell'allegato 5 "Metadati" alle presenti Linee guida.

[Linee guida AgID, All. 1 «Glossario dei termini e degli acronimi»]

Piano di organizzazione delle aggregazioni documentali:

«Strumento integrato con il sistema di classificazione a partire dai livelli gerarchici inferiori di quest'ultimo e finalizzato a individuare le tipologie di aggregazioni documentali (tipologie di serie e tipologie di fascicoli) che devono essere prodotte e gestite in rapporto ai procedimenti e attività in cui si declinano le funzioni svolte dall'ente».



Ministero dell'istruzione e del merito
Ufficio Scolastico Regionale per la Lombardia
Direzione Generale
Via Polesine, 13 – 20139 Milano – Codice Ipa: m_pi

MANUALE DI GESTIONE DOCUMENTALE DI AOO

Piano di Fascicolazione



Versione 1.0
Data pubblicazione: 10/11/2023

La domanda di ricerca

Esempio n. 1

In molte amministrazioni pubbliche e aziende private i **documenti non sono né classificati né aggregati**.

In altri casi, le aggregazioni documentali **non sono ben create**, con il risultato di un numero incontrollato di documenti **non ordinati**, non collocati nella cartella corretta e **difficili da trovare**.

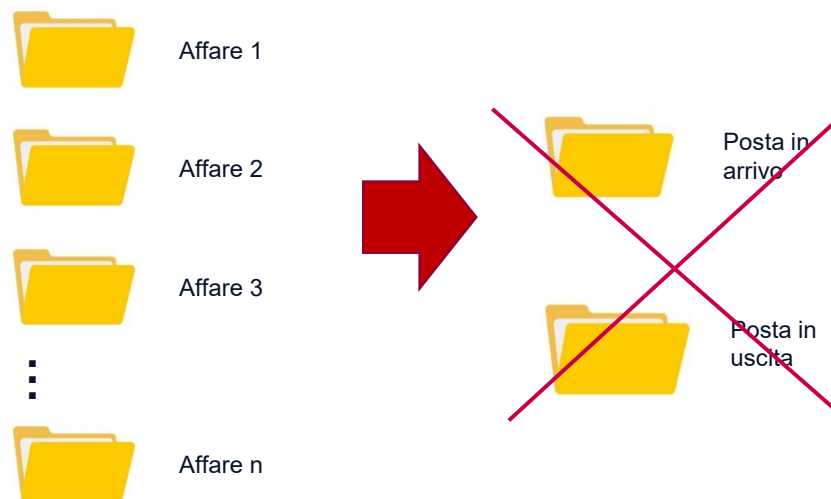
In molti casi **mancano i metadati**, necessari per garantire l'affidabilità, l'attendibilità, la qualità e la sostenibilità delle valutazioni e delle acquisizioni.

Nonostante i progressi compiuti nelle varie tecnologie a supporto della gestione dei documenti, il supporto software per queste attività rimane limitato.



Esempio n. 2

La gestione della posta elettronica è diventata **una delle attività che richiedono più tempo** sia nel settore pubblico che nelle aziende private e nelle attività personali. Le e-mail sono spesso gestite come singoli record senza alcun legame con altre e-mail e **non sono classificate né organizzate in aggregazioni archivistiche (cartelle)**; inoltre non sono collegate e classificate nel sistema di gestione documentale del soggetto produttore.



È possibile utilizzare gli **strumenti di IA** per **classificare** i documenti e creare (o ricreare) le **aggregazioni archivistiche** e individuandone i **metadati**?

2. La ricerca condotta sull'impiego dell'AI per la classificazione e fascicolazione

Il Progetto InterPARES Trust AI e il gruppo di lavoro CU05



InterPARES Trust AI (2021-2026) is a multi-national interdisciplinary project aiming to design, develop, and leverage Artificial Intelligence to support the ongoing availability and accessibility of trustworthy public records by forming a sustainable, ongoing partnership producing original research, training students and other highly qualified personnel (HQP), and generating a virtuous circle between academia, archival institutions, government records professionals, and industry, a feedback loop reinforcing the knowledge and capabilities of each party.

The **I Trust AI** goals are to:

1. Identify specific AI technologies that can address critical records and archives challenges;
2. Determine the benefits and risks of using AI technologies on records and archives;
3. Ensure that archival concepts and principles inform the development of responsible AI; and
4. Validate outcomes from Objective 3 through case studies and demonstrations.

InterPARES Trust AI (2021-2026) è un progetto **interdisciplinare multinazionale** che mira a:

- identificare **tecnologie specifiche di intelligenza artificiale** in grado di affrontare le sfide critiche relative ai documenti e agli archivi;
- determinare i **benefici e i rischi dell'uso delle tecnologie AI su documenti e archivi**;
- garantire che i **concetti e i principi archivistici** siano tenuti in conto nello sviluppo di un'IA responsabile;
- convalidare i risultati attraverso **studi di caso**.

<https://interparestrustai.org/>

The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas

CU05

Stefano Allegrezza (co-chair) (**Università di Macerata**)
Mariella Guercio (co-chair) (Associazione nazionale archivistica italiana - **ANAI**)
Massimiliano Grandi (Associazione nazionale archivistica italiana - **ANAI**)
Bruna La Sorda (Associazione nazionale archivistica italiana - **ANAI**)
Francesca Magnoni (North Atlantic Treaty Organization - **NATO**)
Maria Mata Caravaca (International Centre for the Study of the Preservation and Restoration of Cultural Property - **ICCROM**)
Samir Musa (Historical Archives of European Union – **HAEU**)
Luís-Esteve Casellas Serra, **Municipality of Girona – Spain** (connection with AA01 “Employing AI for Retention & Disposition in Digital Information and Recordkeeping Systems (DIRS)”)
Nicola Di Matteo (**Halifax University**, Canada)
Eleonora Luzi (Associazione nazionale archivistica italiana - **ANAI**)
Gianni Penzo Doria (**Università dell’Insubria**)
Gabriele Bezzi (Associazione nazionale archivistica italiana - **ANAI**)
Stefano Delli Ponti, **Roberto D’Ippolito** (SIAV)
Marianna Tascone (**PAR-ER, Regione Emilia-Romagna**)
Gabriele Bezzi (**ANAI Emilia Romagna**)

La domanda di ricerca

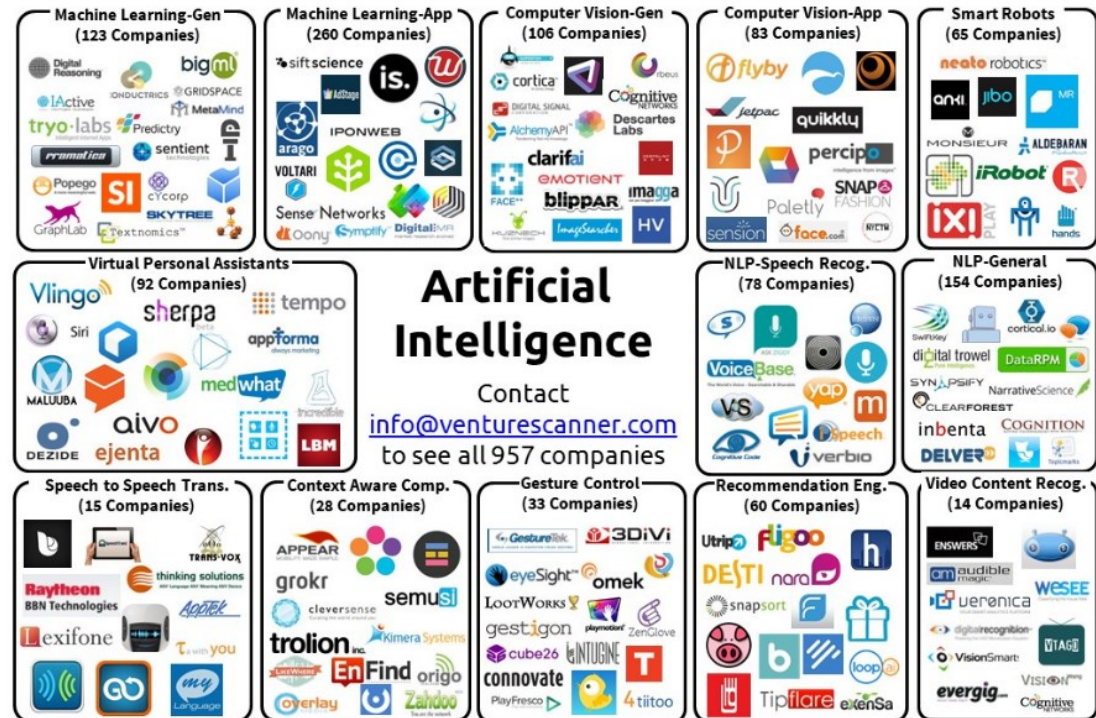
È possibile utilizzare gli **strumenti di IA** al fine di **classificare** i documenti, creare (o ricreare) le **aggregazioni documentali**, l'individuazione dei **metadati**?



Quali **tecnologie AI** potrebbero essere utili per la gestione automatica o semi-automatica di operazioni archivistiche quali:

- la **classificazione** automatica?
- la creazione delle **aggregazioni documentali**?
- l' **integrazione** dei metadati?
- per **filtrare** e **aggregare** le e-mail?

Ci sono **migliaia di aziende** che dichiarano di utilizzare l'IA. Centinaia di loro dichiarano di utilizzare le tecnologie AI sul campo o gli ERMS/EDMS.



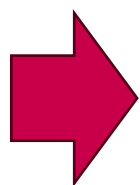
Come verificare le funzionalità degli strumenti proposti da queste aziende e come verificare se questi strumenti funzionano davvero?

Fase 1: Identificazione delle aziende di IA

Identificazione di un gruppo iniziale di **300 aziende** interessanti per lo studio, ovvero aziende che sviluppano Soluzioni basate su tecnologie legate all'IA e sono pertinenti all'ambito dello studio CU05.

Strumenti per la costruzione dell'elenco:

- **ricerche dirette su Internet** utilizzando parole chiave e stringhe di testo;
- **le risorse e le conoscenze messe a disposizione dai professionisti** (Alan Pelz-Sharpe, Andrew Warland, James Lappin, Jenny Bunn e Paul Young)



L'elenco non è né esaustivo né definitivo, ma è un punto di partenza.

LIST OF 100 COMPANIES THAT ARE OR MIGHT BE RELEVANT TO CU05

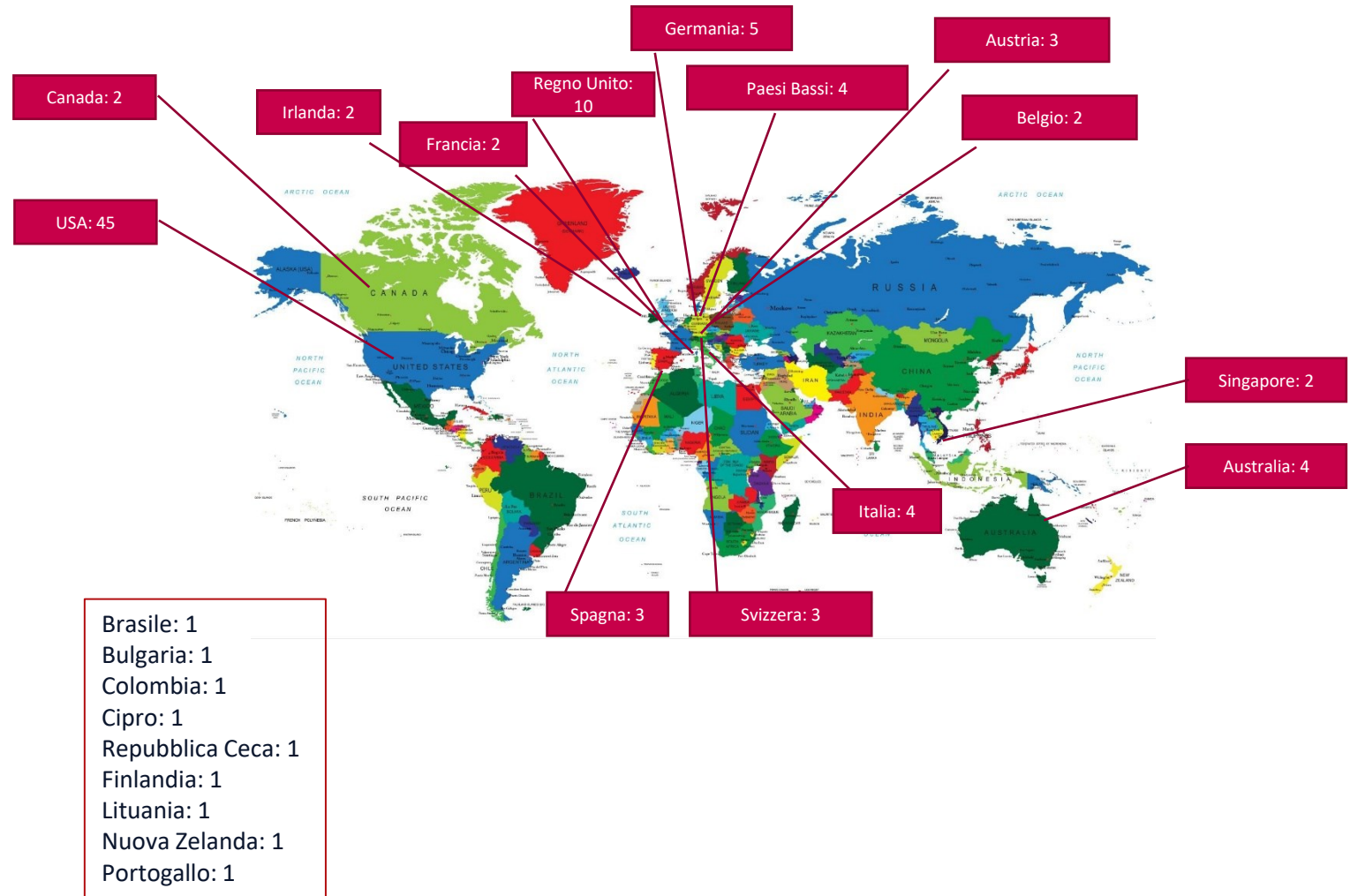
Nr.	Company	Location	Website	Rating	Notes
1	a3doc Wolters Kluwer	Alphen aan den Rijn, The Netherlands	http://www.a3doc.com/experiencia-clientes-gestion-documental-cloud-colaborativa.html	1	a3doc Wolters Kluwer has developed a3doc cloud, an application for document management which is able to classify and store automatically corporate documents. There is no much information in the webpage describing the product. The company is based in the Netherlands, but the webpage is in Spanish.
2	Abbyy	California, United States	https://www.abbyy.com/	1	The main headquarters of Abbyy are Milpitas, California, USA. Abbyy mainly develops application for data capture, but one of their software products is ABBYY Vantage, software for Intelligent Document Processing. ABBYY Vantage uses AI-driven technologies to process "documents of any kind—structured, semi-structured, or unstructured, and all type of data including machine printed, hand printed, barcodes, signatures, and check boxes". In ABBYY Vantage "Trained skills can be quickly designed to understand and extract information from all types of documents". "Once skills are deployed, Vantage then monitors, measures, and analyses performance of all your deployed skills, creating new learning models—so you can continuously improve and move your automation to the next level." ABBYY "Vantage skills are continuously getting smarter and more accurate over time, as new document variations and statistical data is collected during human-in-the-loop review". "The easy-to-use, no-code Vantage platform can be utilized to set up and train Document, Classification, and Process Skills for just about any document type and flow" - For this information see https://www.abbyy.com/vantage/
3	ActiveNav	Reston, Virginia, USA	activenav.com	1	ActiveNav is based in US, UK and Australia. They seem to deal mainly with automated data mapping and automated data classification. As it is the case for Automated Intelligence, you do not find a specific mention of records and archives, but clearly the services they offer (or at least say they can offer) also concern archives and records management.
4	Acumatica - Webiplex - PairSoft	Washington, USA (Acumatica) Florida, USA (PairSoft)	https://www.acumatica.com/media/2017/04/DocuPeak-for-Acumatica.pdf	1	DocuPeak is an application developed by Webiplex - that has been recently purchased by PairSoft, a company based in Florida, USA. DocuPeak has been built to be integrated with the Enterprise Resource Planning platform created by Acumatica, whose main headquarters is in Seattle, Washington, USA. DocuPeak is powered by Robotic Process Automation technologies. This is the introductory description of DocuPeak: "Robotic Process Apps built on the DocuPeak cloud platform streamline operations 'before and including data entry', from automated data extraction and data entry, to approval workflows and document lifecycle management to entirely electronic forms-based applications". Some features of DocuPeak are "Leverage Smart Document Recognition (SDR) to extract key data from documents, without templates, automating document indexing and transactional data entry"; "DocuPeak's Rules Engine to automatically route documents such as AP invoices through a predefined role based review and approval process, including notifications and escalation procedures"; "Integrate all documents to the associated transaction within Acumatica for synchronization of key document data and instant retrieval"; "Create a secure, compliant document management environment, maintaining an audit trail of all document-based activity, including check-in, check-out and version control on changed documents".
5	Ademero	Florida, United States	https://www.ademero.com/	1	The company is based in Lakeland, Florida, USA. Ademero is a Document Management Software platform. Thanks to its AI capabilities, Ademero can "intelligently identify, categorize and process accounts payable invoices or any other low or high volume paper entering Content Central" - see https://www.ademero.com/document-management-software/ . Ademero includes 2 modules, one for document scanning / capture, and another one for document management after the capture of the document. "Capture Software lays the foundation for an automated onboarding process by both classifying and indexing documents, then working hand-in-hand with a Document Management System and your other business software solutions, streamlines document and office workflows. "Intelligent" or fully-automatic solutions in this category means that your staff simply scan in paper documents at your office multifunction printer (mfp) and the software uses optical character recognition (OCR) to turn that scanned image into a digital version of that document, then handles classifying and indexing each document before handing it off to your DMS or other software applications." As to the Document Management module, one of its most interesting features is "the logical and consistent folder and file building it provides". "You just upload your document and the system will handle filing it away so that you or anyone who has permission to access those documents can locate it quickly by navigating logically named folders based on the standards you require".
6	Adlib	Burlington, Ontario, Canada	www.adlibsoftware.com	3	Adlib is a Canadian company which has produced "Adlib Elevate". "Adlib Elevate" is a File Analytics platform to automate discovery, extraction and classification of vital data from complex documents to streamline data-intensive processes and accelerate process automation. Adlib Elevate is one of the 5 suppliers AI-based application for records management assessed by The UK National Archives.
7	AIDA Cloud	Turin, Italy	https://www.aidacloud.com/home	0	AIDA has been developed by Technology & Cognition LAB, based in Turin, Italy. AIDA is a document retrieval application that uses AI "to recognise any type of document and, depending on the user's needs, extract all the information needed, with a simple learning process that requires no technical knowledge". Basically it retrieves the information you need, organizes it and makes it available for users - see https://doc.aidacloud.com/aida/ Users defining document types, and AIDA through Intelligent Document Analysis manages to recognise such document

Fase 1: 100 aziende - distribuzione geografica

L'elenco è stato successivamente limitato a **100 aziende**

Le caratteristiche dei loro **software di intelligenza artificiale** sono state analizzate in base alle **informazioni disponibili sui loro siti web** con riferimento a:

- in cui l'azienda dichiara che la **gestione dei documenti** è uno degli obiettivi delle sue applicazioni di intelligenza artificiale;
- **evidenza di interesse** per gli aspetti del **records management** e degli **archives** (anche se in alcuni casi ciò non è dichiarato apertamente ma si può solo intuire dal contenuto del sito web).



Fase 1: Identificazione delle aziende di IA

Poiché non era possibile intervistare tutte le **100 aziende**, dall'elenco iniziale abbiamo selezionato un elenco di **28 aziende** sulla base di:

- il loro **portafoglio**
- il loro coinvolgimento diretto nel **campo della gestione dei documenti**
- la loro **conformità** ai quadri normativi e agli standard rilevanti per il settore
- la **reputazione** generale dell'azienda.

1	Microsoft	Washington, DC, USA	https://www.microsoft.com/en-gb
2	Iron Mountain	Boston, Massachusetts, USA	www.ironmountain.com
3	Adlib	Burlington, Ontario, Canada	www.adlibsoftware.com
4	Castlepoint	Canberra, Australia	www.castlepoint.systems
5	Gimmel	Texas, USA	https://www.gimmel.com/
6	Quest-it	Siena, Italia	www.quest-it.com
7	Gruppo Adapting	Valencia, Spagna	https://www.adapting.com/en/
8	Hyland	Westlake, Ohio, USA	https://www.hyland.com/en
9	Stratagemma	Aurora, Colorado, USA	www.stratagemgroup.com
10	Aluma	Cambridge, Regno Unito e New York, USA	https://aluma.io/
11	Collabware	Washington, DC, USA	collabware.com
12	Ephesoft	Irvine, California, USA	https://ephesoft.com/
13	Leggi-Coop	Innsbruck, Austria	https://readcoop.eu/transkribus/
14	Punto di registrazione	Sydney, Australia	www.recordpoint.com
15	Software Prisma	California, USA	https://prismsoftware.com/
16	Sistema esperto	Modena, Italia	https://www.expert.ai/
17	GRMGestione dei documenti	New Jersey, USA	https://www.grmdocumentmanagement.com/
18	Grooper	Oklahoma, USA	https://www.bisok.com/intelligent-document-processing/
19	Ripcord	Hayward, California, USA	www.ripcord.com
20	Cortical	New York, USA	www.cortical.io
21	AmyGB.ai	Mumbai, India	www.amygb.ai
22	Bizamica	Pune, India	www.bizamica.com
23	Docxflow	Popayán, Colombia	https://www.docxflow.com/
24	IA Gleematico	Singapore	https://gleematic.com/
25	Soluzioni aziendali SBK	São Bernardo do Campo, San Paolo, Brasile	www.sbkbs.com.br
26	Datacentrix	Johannesburg, Sudafrica	www.datacentrix.co.za



Anyz (Norvegia)
DXC (Italia)

Fase 2. Questionario e interviste

Per raccogliere informazioni più precise, abbiamo preparato un **questionario** molto dettagliato, volto a raccogliere sistematicamente le informazioni per un'adeguata valutazione delle domande.

Abbiamo inviato alle **28 aziende** una **lettera di invito** ufficiale (in inglese, spagnolo o portoghese, a seconda della lingua preferita dall'azienda) a partecipare all'indagine.

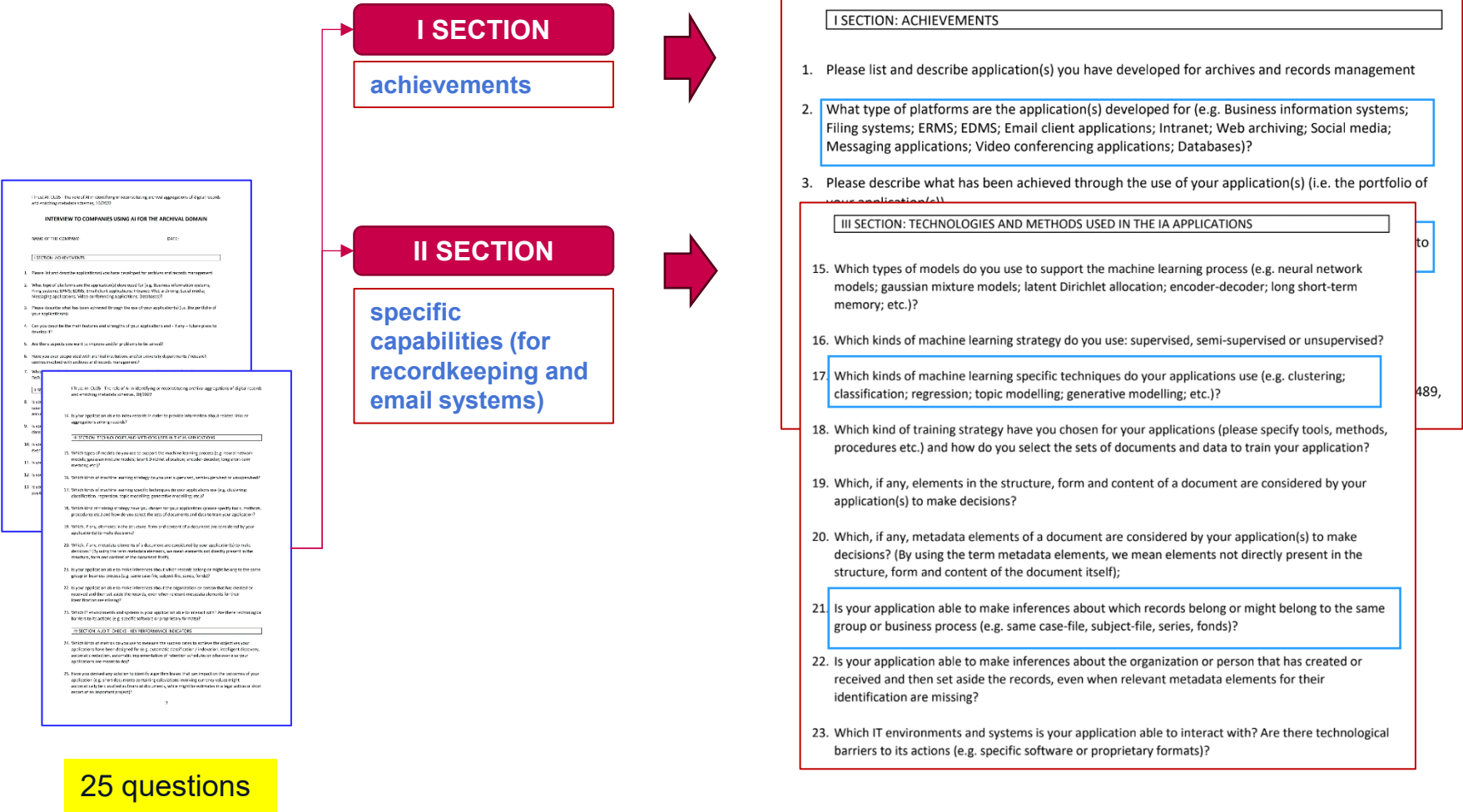
Il questionario è stato spiegato oralmente durante un **incontro preliminare** con il personale di gestione delle informazioni e gli ingegneri informatici.

Successivamente, le aziende hanno compilato il questionario disponibile su **Google Forms**.

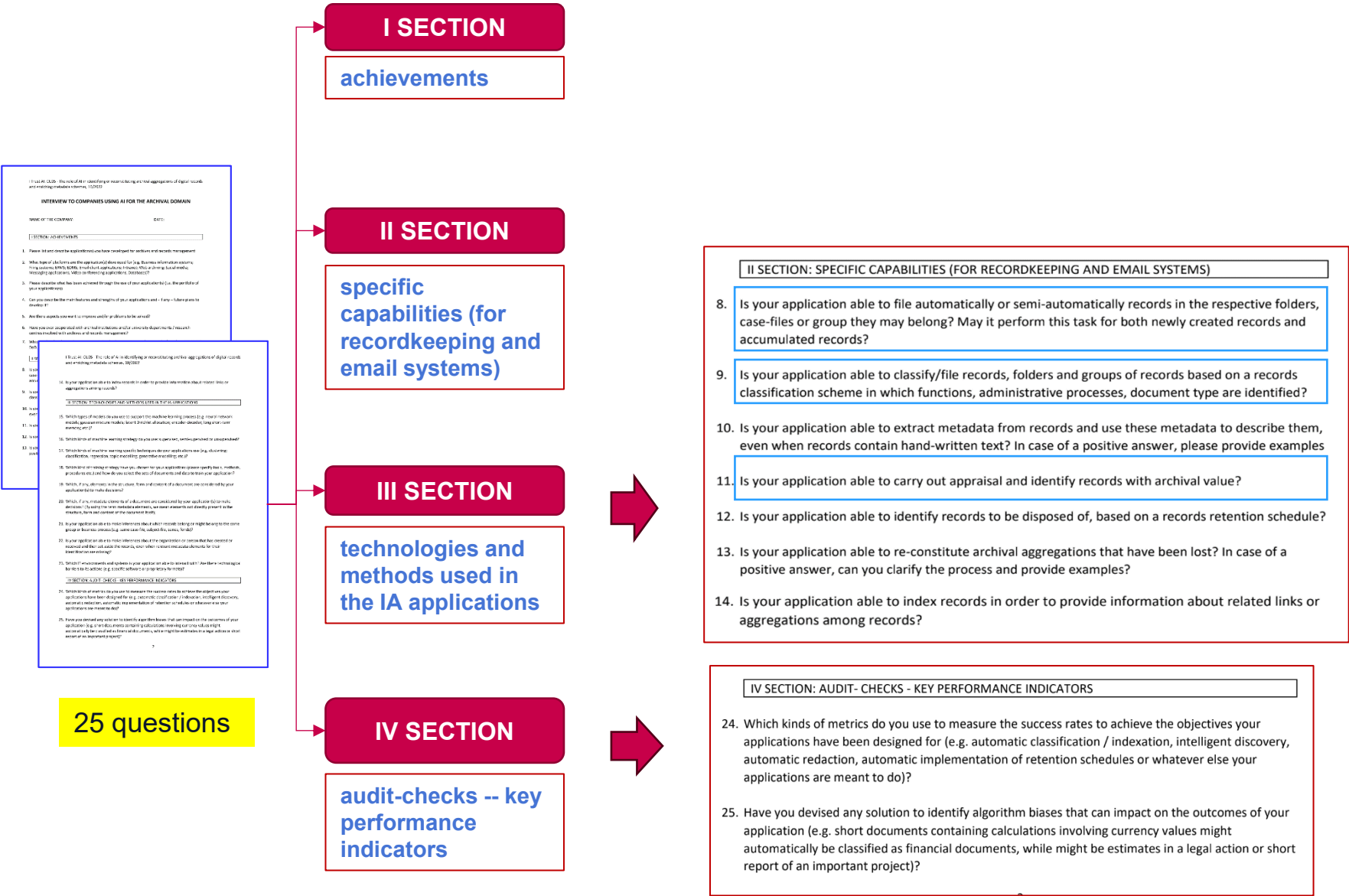
The image shows a screenshot of a Google Form titled 'INTERVIEW TO COMPANIES USING AI FOR THE ARCHIVAL DOMAIN'. The form is part of the InterPARES Trust AI project. It includes a section for 'Introduction to the survey' with the following details: Project: InterPARES Trust AI (ITrustAI), Working Group: Creation and Use (WG1), Study: The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas, and Study code: CU05. There is a 'Research Questions' section asking if AI tools can be used to constitute or reconstitute archival aggregations and create metadata schemas for them. A 'Description' section explains the problem of missing metadata, necessary to ensure the reliability, trustworthiness, quality and sustainability of appraisal and acquisition, is common and complex today. It concerns the uncontrolled creation of a huge amount of records in the active phase. Despite progress on various technologies to support record management, software support for those activities remains limited. To identify possible concrete areas where AI technologies could play a crucial role, the first fundamental step is to identify the specific and most common scenarios we face in the digital dimension, such as: 1. Records managed by ERMS without the full set of information required for proper records creation; 2. Business systems that create and manage records with only partial identification and procedural information; 3. Records created by systems without metadata and without being integrated in ERMS, including email repositories.

https://docs.google.com/forms/d/e/1FAIpQLSc8US3a89JbVjhfdma2EqYm1Xo_LVqP3qh_7kM4CJpKQStTg/viewform

Fase 2. Questionario e interviste

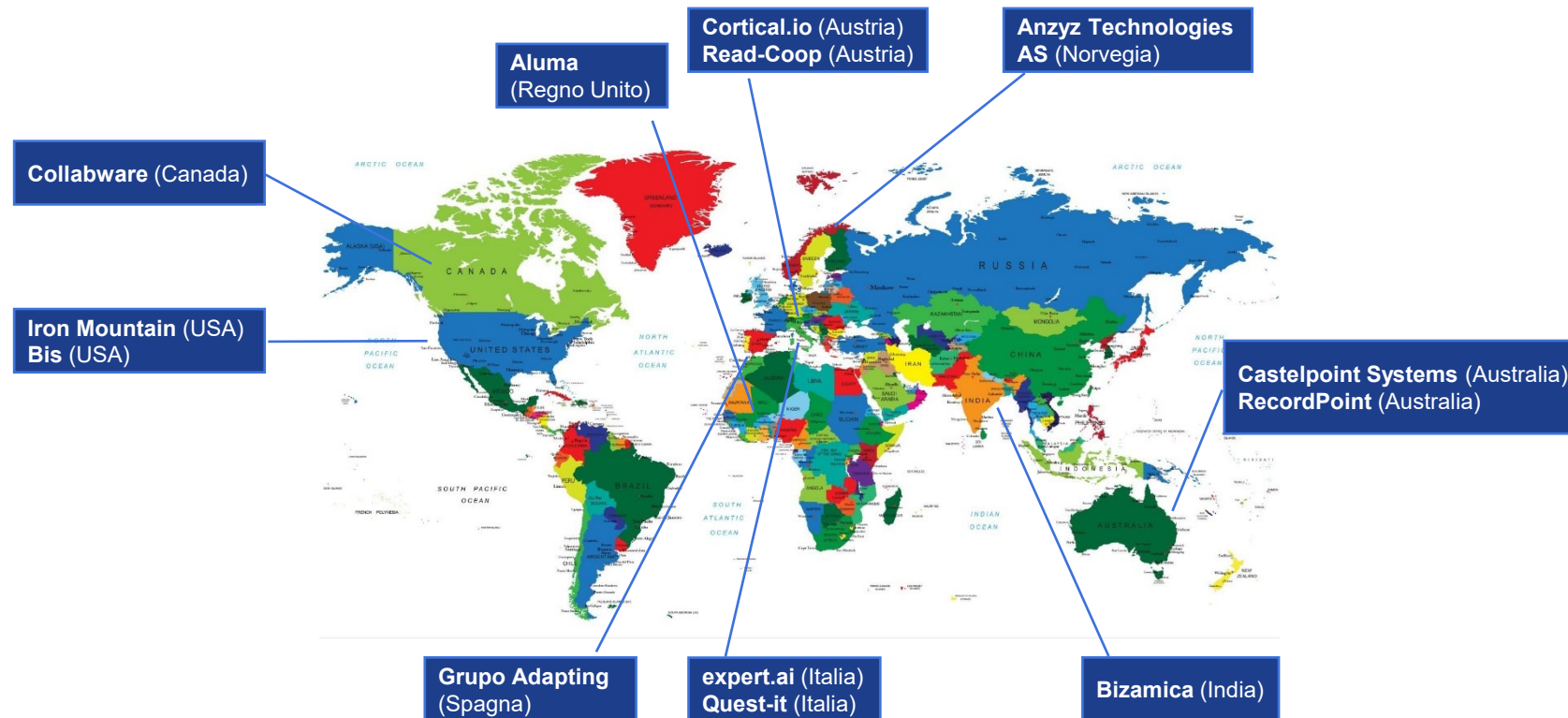


Fase 2. Questionario e interviste



Fase 2. Aziende che hanno risposto al sondaggio

Hanno risposto **13 aziende**:



Il rapporto finale



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	10.0
Writers	Stefano Allegrezza, Mariella Guercio, Maria Mata Caravaca, Massimiliano Grandi, Bruna La Sorda
Date	November 1, 2023

Stefano Allegrezza (co-presidente) (Università di Bologna - Istituto di Ricerca per l'Intelligenza Artificiale Centrata sull'Uomo-ALMA AI)



Mariella Guercio (co-presidente) (Associazione nazionale archivistica italiana - ANAI)



Maria Mata Caravaca (Centro Internazionale per lo Studio della Conservazione e del Restauro dei Beni Culturali - ICCROM)



Massimiliano Grandi (Associazione nazionale archivistica italiana - ANAI)



Bruna La Sorda (Associazione nazionale archivistica italiana - ANAI)



3. Risultati

Alcuni risultati della survey

Automatic classification

- analysis of **metadata elements** available both in the records and aggregations (case-folder specifications)
- identification of **document type**
- in case the available metadata should prove to be insufficient for classification, then classification is based on the **record content**
- generation of **labels and tags** belonging to any record classification scheme (taxonomy or term ontology)

Yes: 10

No: 3

Inferences on records grouping

- based on **content and/or context**
- if there is **metadata** to represent those processes (e.g. a case file number)

Yes: 9

No: 4

Indexing to create links, aggregations

- difficult task **if contextual data is lost**

Yes: 6

No: 7

Filing / Aggregating

- by **document type**
- by considering the original structure of the content source
- generation of **labels and tags** from any record classification scheme, based on the **record content**

Yes: 10

No: 3

Inference on organisation or person

- If the involved entities are stated in the **content** of the document

Yes: 7

No: 6

Re-establish the archival bond of unarranged records

- extraction of data **from records content or from metadata**

Yes: 5

No: 8

Soluzioni tecnologiche

Techniques and Analysis Models

The companies listed a wide range of different **analysis models**: **24 different entries** - the 5 most recurring are:

Neural Network Models (4 companies)

Support Vector Machines (4 companies)

Decision Trees (3 companies)

Random Forests (3 companies)

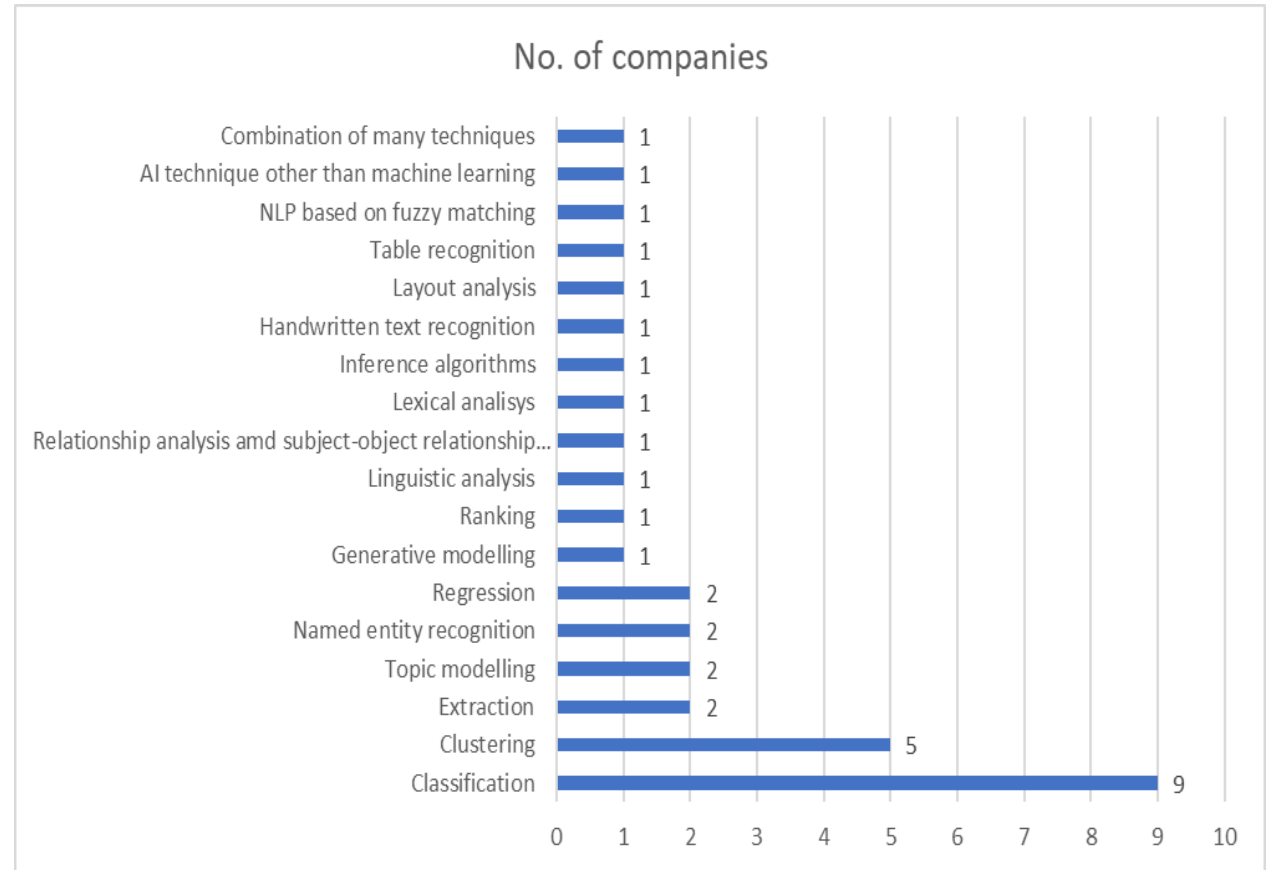
LSTM - Long-Short Term Memory (3 companies)

As to **the types of techniques** featured in the products of the companies - the 2 most recurring are:

Classification (9 companies)

Clustering (5 companies)

This is no surprising as the companies were selected because of their expertise at least in document management. However **18 different types of techniques** overall have been reported

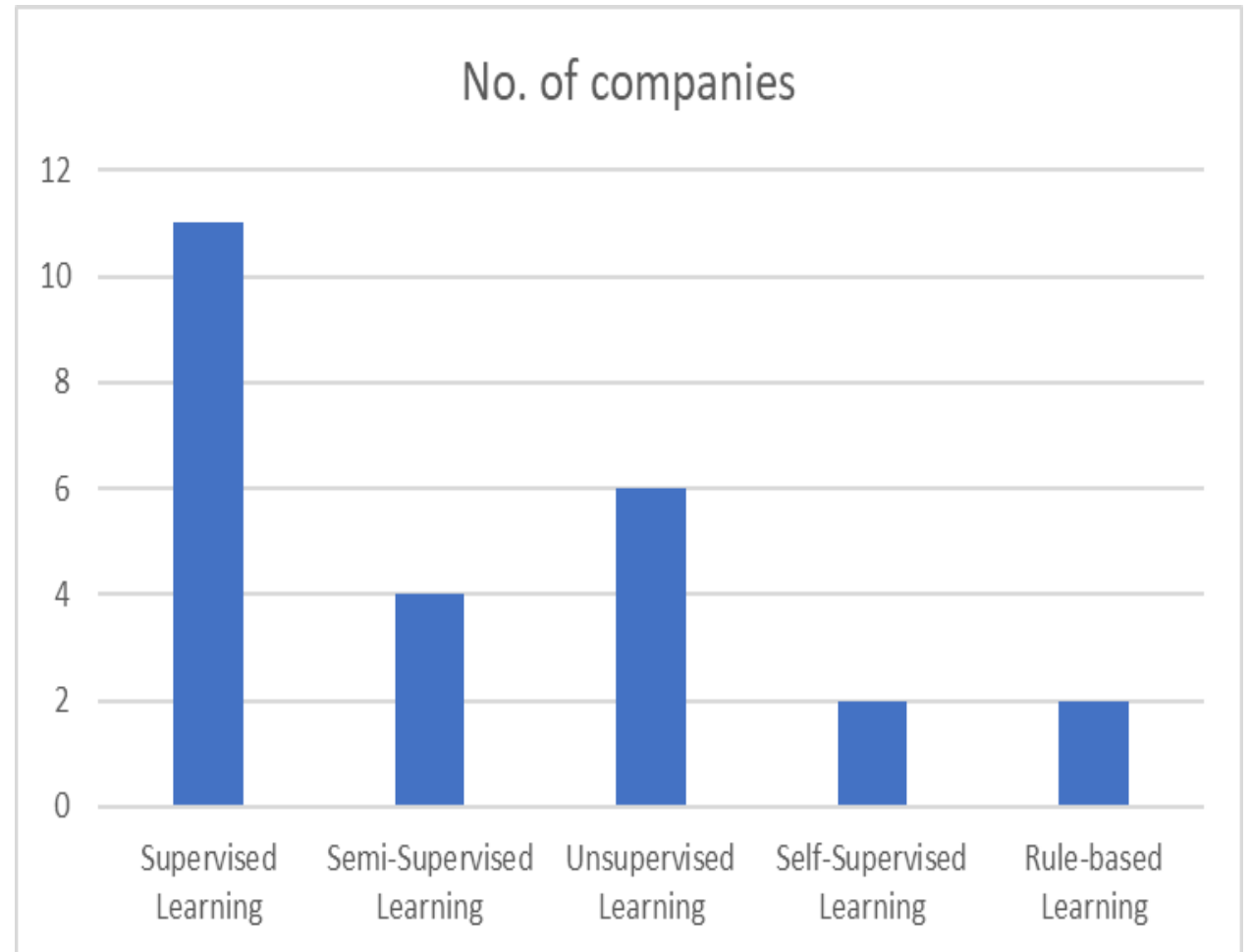


Soluzioni tecnologiche

Training strategies

The companies declared they use a combinations of different training strategies:

- **Supervised Learning**: 11 companies
- **Semi-Supervised Learning**: 4 companies
- **Unsupervised Learning**: 6 companies
- **Self-Supervised Learning** : 2 companies
- **Rule-based Learning** : 2 companies



Involvement with records management and archives

Le aziende dichiarano di essere conformi (o sono state progettate per supportarle) alle seguenti norme:

- **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);

Risultati dell'indagine dal punto di vista archivistico

Per quanto riguarda la **classificazione**, tutte le aziende intervistate hanno dichiarato di:

- aver sviluppato soluzioni basate su tecnologie di intelligenza artificiale per l'**indicizzazione** e/o la **classificazione** di documenti/dati strutturati, semi-strutturati e non strutturati,
- basate su **tecniche di apprendimento automatico** e sull'**estrazione automatica dei dati**.

Il ruolo dei **metadati presenti o dedotti** è sempre al centro di tutte le risposte.

L'informazione sulla **tipologia dei documenti** - quando disponibile - è spesso considerata un'altra componente cruciale per il successo dell'applicazione delle tecniche di IA ai documenti.



Risultati dell'indagine dal punto di vista archivistico

Per quanto riguarda la classificazione dei documenti:

- solo un'azienda ha dichiarato che la sua piattaforma **può essere addestrata**, grazie a uno specifico set di dati, per generare autonomamente (senza intervento umano) etichette e tag relativi a qualsiasi **schema di classificazione**,
- in tutti gli altri casi **l'intermediazione umana è considerata insostituibile** per fornire risultati coerenti.



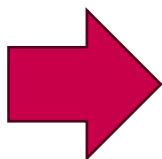
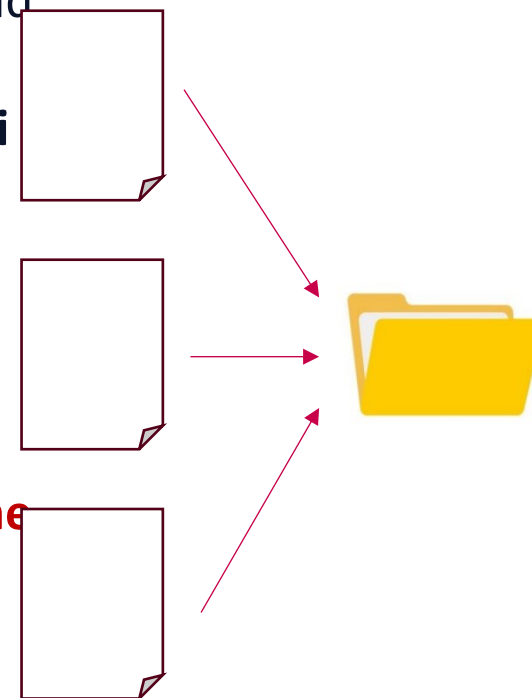
Risultati dell'indagine dal punto di vista archivistico

Per quanto riguarda la creazione delle **aggregazioni documentali** (o la ri-costituzione), le promesse di automatizzazione non sono molto incoraggianti, poiché si conferma che questa possibilità è **limitata a casi molto specifici**, come ad esempio:

- quando esistono delle **specifiche ben definite sulle tipologie di documenti**
- quando vi sono informazioni sulla **struttura originale** e ciò consente **l'individuazione di relazioni funzionali tra i documenti**

L'aggregazione **automatica** o **semi-automatica** basata sul contenuto del documento è **solo suggerita** e di solito deve comunque essere **supportata dalla convalida da parte dell'utente** o da **regole disponibili al momento della creazione**

Anche la **ri-costituzione del vincolo archivistico** - quando perduto o non esplicitamente definito - è riconosciuta come un'**attività complessa**, difficilmente realizzabile senza l'aiuto significativo da parte dell'uomo e/o informazioni descrittive.



Nella maggior parte dei casi le funzionalità **non sono già completamente sviluppate**, ma **in via di sviluppo** e ciò implica **maggiori investimenti (sostenuti dal mercato?)**.

Risultati dell'indagine dal punto di vista archivistico

La terminologia è una sfida cruciale.

Quando si interagisce con le aziende, la comunità archivistica deve prestare molta attenzione:

- chiarire i concetti di termini generali come la differenza tra **documento** e **dato**, il concetto di **aggregazione**, la differenza tra **classificazione, indicizzazione e categorizzazione**;
- interpretare correttamente espressioni come "**Intelligent Document Processing**" che, di solito, non ha nulla a che fare con il documento, il suo trattamento e, in ultima analisi, con l'**intelligenza archivistica**.



4. Considerazioni finali

Considerazioni finali

La maggior parte delle aziende intervistate ha dimostrato:

- di **comprendere la complessità dell'ambiente e delle funzioni archivistiche**
- di essere consapevoli dell'**importanza dei metadati originali acquisiti nelle attività correnti del soggetto produttore**, sia nel caso in cui si tratti di classificare automaticamente i documenti sia nel caso si tratti di creare le aggregazioni documentali

L'indagine testimonia che la **complessità delle funzioni archivistiche** non può essere **facilmente ridotta** e rimossa da un approccio automatico, ma solo **supportata** dalle tecnologie AI attraverso **l'intermediazione di archivisti ed utenti**.



Future work



**Casi di
studio**



NATO (North Atlantic
Treaty Organization)



ParER

Polo archivistico dell'Emilia-Romagna

PAR-ER (Polo archivistico
Regione Emilia Romagna)



CISU (Centro Italiano
di Studi Ufologici)

Considerazioni finali

- Dalla intelligenza artificiale, essendo basata su **modelli probabilistici** e non deterministici (come gli expert systems degli anni 80-90), non ci si può aspettare che classifichi i documenti **in maniera corretta al 100%** (esattamente come non ci si può aspettare che individui sempre correttamente l'immagine di un gatto tra le tante che le vengono sottoposte). Quindi gli errori (**hallucinations**) ci saranno sempre.
- Questo, però, è esattamente quello che avviene con gli esseri umani: anche gli esseri umani commettono degli errori. Ricordiamoci che l'uomo non è una macchina e quindi a volte sbaglia!
- Quello che ci si può ragionevolmente aspettare è che migliorando l'addestramento dell'AI (ad es. aumentando la dimensione dei corpora documentari) il numero di errori diminuisca sempre più e migliori sempre più l'accuratezza.
- Al momento la verifica umana sembra necessaria, ma sarà sempre così?
- Sì, fino a quando la precisione dell'intelligenza artificiale non sarà comparabile o addirittura superiore a quella dell'essere umano (e in taluni campi già è così) ed allora davvero le operazioni di **classificazione** e di creazione delle **aggregazioni documentali** potranno essere svolte integralmente dall'intelligenza artificiale senza intervento umano.

Grazie per l'attenzione!

Stefano Allegrezza

Università di Macerata (ITALIA), Dipartimento di Studi umanistici

stefano.allegrezza@unimc.it