Webinar

L'intelligenza artificiale negli archivi: soluzioni in campo e casi concreti

10 giugno 2025 - ore 11:30

Termine iscrizioni: **09 giugno 2025 ore 14:00**

L'intelligenza artificiale è sempre più presente anche nel dibattito sugli archivi e sulla gestione documentale. Ma cosa è davvero realtà e cosa narrazione?

Il webinar vuole offrire un momento di riflessione sulla crescente crucialità dell'intelligenza artificiale nel governo delle fonti, soprattutto in relazione alla garanzia della loro autenticità e affidabilità, ruolo questo, da sempre affidato all'archivista.

L'obiettivo è fornire, per quanto possibile, uno sguardo critico e aggiornato sull'impiego concreto dell'IA nel settore archivistico, un'occasione per distinguere tra narrazione e pratica e per capire dove stia andando davvero l'innovazione negli archivi.

L'evento sarà inoltre occasione per offrire una panoramica su alcune applicazioni già operative e su soluzioni offerte dal mercato, sia in Italia che all'estero, nonché per raccontare un caso di applicazione pratica, in cui l'IA è stata impiegata per la valutazione dei formati dei file, in particolare elaborando l'indice di interoperabilità descritto nell'Allegato 2 alle Linee guida sulla formazione, gestione e conservazione del documento informatico di Agid.

Il webinar si inserisce all'interno di un ciclo previsto dall'Accordo di collaborazione sulla conservazione tra **Formez PA e Regione Emilia-Romagna (Parer).**

Il termine per le iscrizioni al webinar è **lunedì 9 giugno 2025 – 14:00, salvo esaurimento posti.**

Programma

11.30 - 11.35 Introduzione (**Alessandra Cornero**, Formez PA)

11.35 - 11.55 Intelligenza artificiale e archivi: a che punto siamo, oggi (**Giorgia Di Marcantonio**, Università degli Studi di Napoli, l'Orientale)

11.55 - 12.15 Classificare e fascicolare i documenti con l'AI: l'esperienza del progetto InterPares Trust (**Stefano Allegrezza**, Università degli Studi di Macerata)

12.15 - 12.35 Calcolo automatico dell'indice di interoperabilità dei formati (**Marianna Tascone,** Polo archivistico e gestione documentale della Regione Emilia-Romagna- ParER)

12.35 - 13.00 Domande e risposte

Partecipanti

857 iscritti a questo evento

Contatti

ufficio.competenzadocumentale@formez.it

Progetto di riferimento

Attività istituzionali

Ulteriori informazioni

Argomento: *Conservazione, Gestione documentale, Intelligenza artificiale*

Politica di registrazione: *Senza approvazione*

Rimani aggiornato

Accedi per rimanere aggiornato

unimc UNIVERSITÀ DI MACERATA
l'umanesimo che innova

InterPARES Trust AI

anai Associazione Nazionale Archivistica Italiana

Classificare e fascicolare i documenti con l'AI: l'esperienza del progetto InterPares Trust AI

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 nell'ambito del progetto InterPares Trust AI
3. Risultati
4. Considerazioni finali

1. Introduzione

La classificazione dei documenti informatici

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

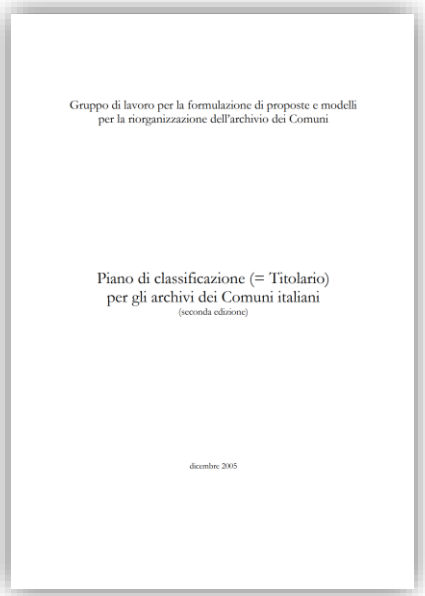
La **classificazione** ha il fine di organizzare logicamente tutti i documenti amministrativi informatici prodotti o ricevuti da un ente nell'esercizio delle sue funzioni. L'attività di classificazione si avvale del **piano di classificazione** che mappa, su più livelli gerarchici, tutte le funzioni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

[Piano di classificazione per gli archivi dei comuni italiani]: «Il piano di classificazione o titolario è il sistema precostituito di partizioni astratte, gerarchicamente ordinate (dal generale al particolare), fissate sulla base dell'analisi delle funzioni dell'ente, al quale deve ricondursi la molteplicità dei documenti prodotti, per organizzarne la sedimentazione ordinata».


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

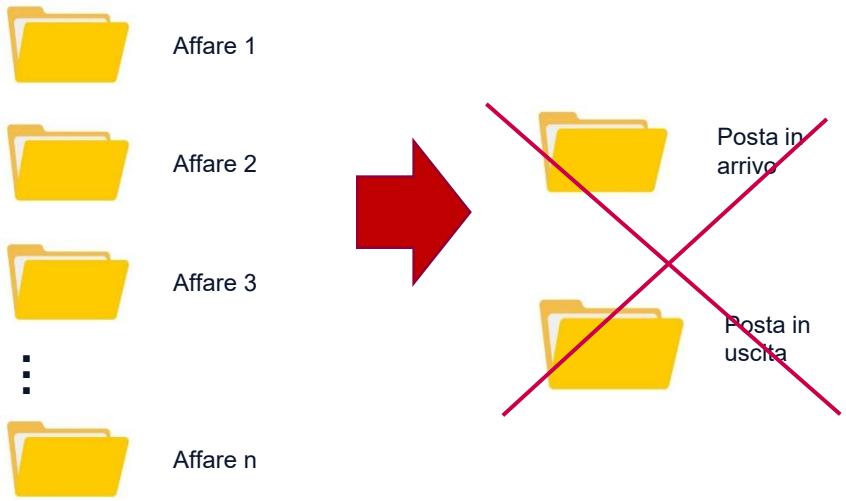
Piano di Fascicolazione

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.

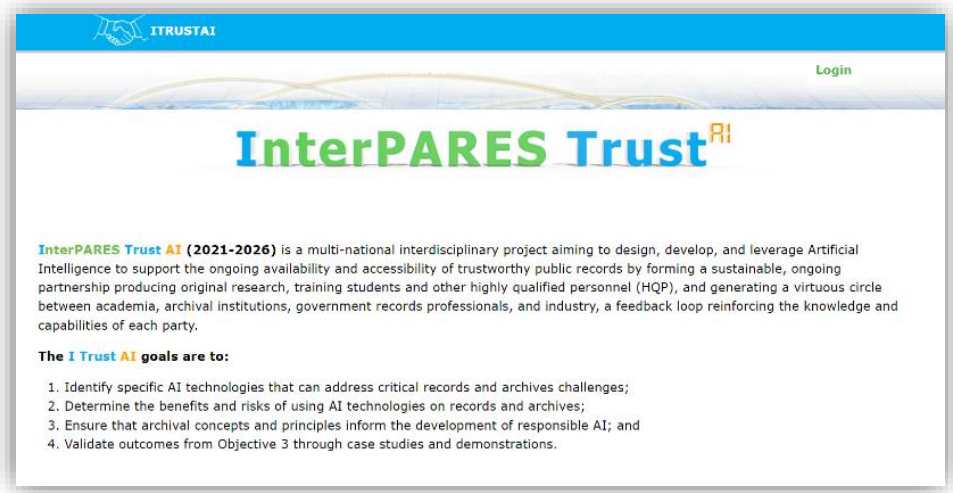
Esempio n. 2
La gestione della posta elettronica è diventata **una delle attività che richiedono più tempo** sia nel settore pubblico che nelle aziende private e nelle attività personali.
Le e-mail sono spesso gestite come singoli record senza alcun legame con altre e-mail e **non sono classificate né oganizzate 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 documentali** e individuare 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) 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**)

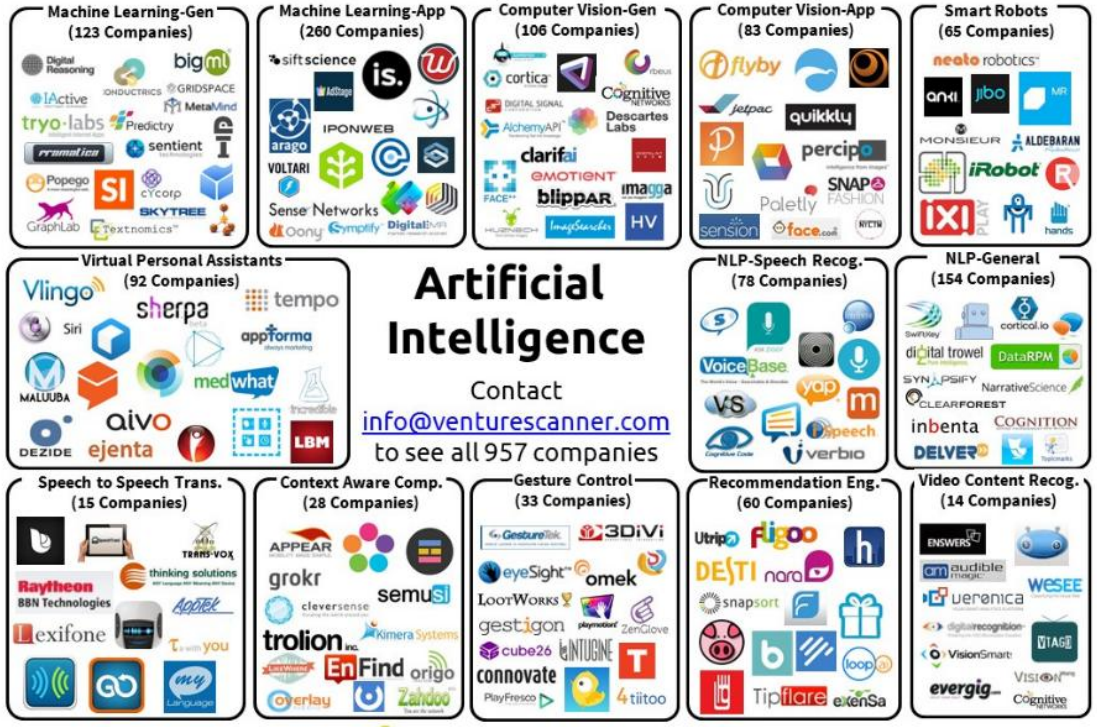
La domanda di ricerca

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 l'esecuzion automatica o semi-automatica di operazioni archivistiche quali:

- la **classificazione dei documenti?**
- la creazione delle **aggregazioni documentali?**
- l' **integrazione dei** metadata?
- la gestione delle **aggregazioni** nel caso della posta elettronica?

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)

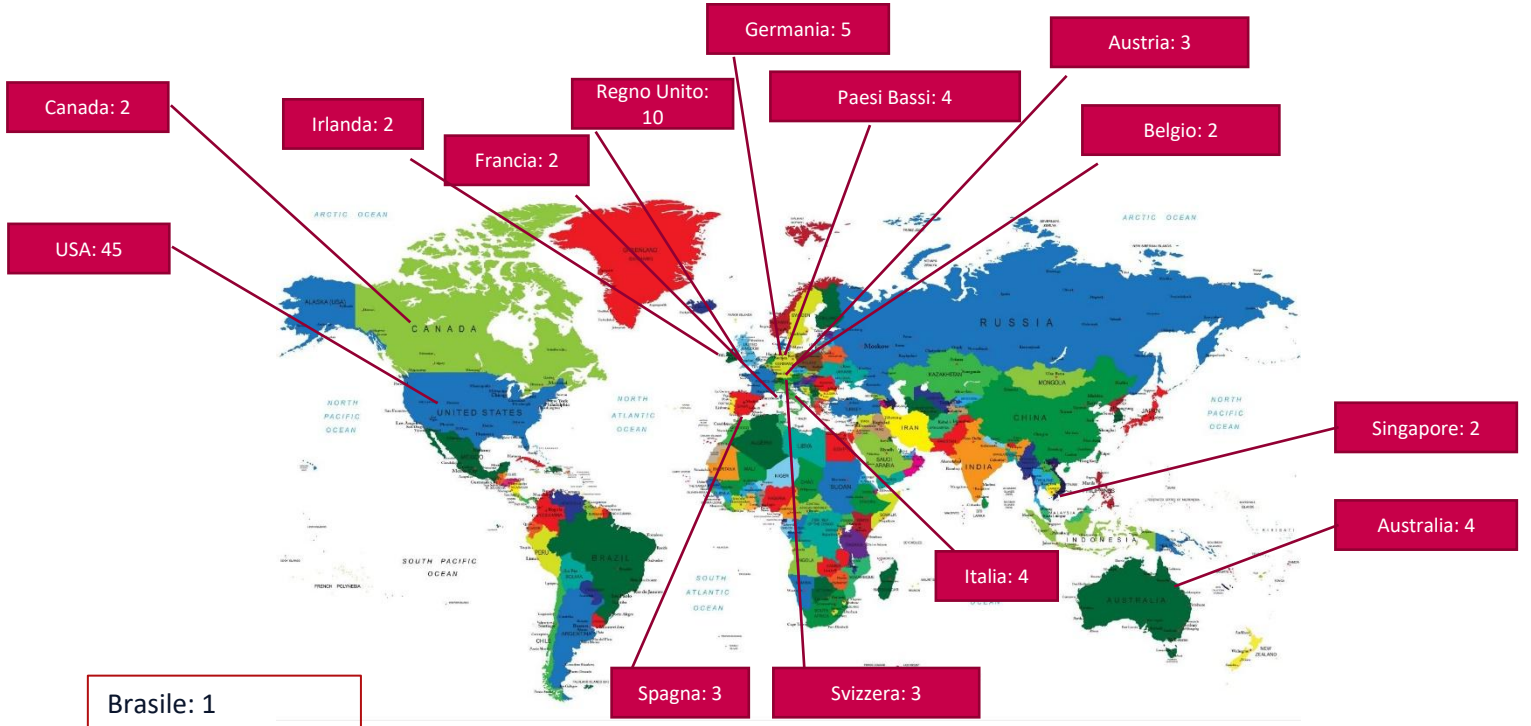
L'elenco non è né esaustivo né definitivo, ma è un punto di partenza.

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 of 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, USAAdemero 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 define 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).



Brasile: 1
Bulgaria: 1
Colombia: 1
Cipro: 1
Repubblica Ceca: 1
Finlandia: 1
Lituania: 1
Nuova Zelanda: 1
Portogallo: 1

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	Gimmal	Texas, USA	https://www.gimmal.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	Corticale	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

\+ Anzyz (Norvegia)
DXC (Italia)

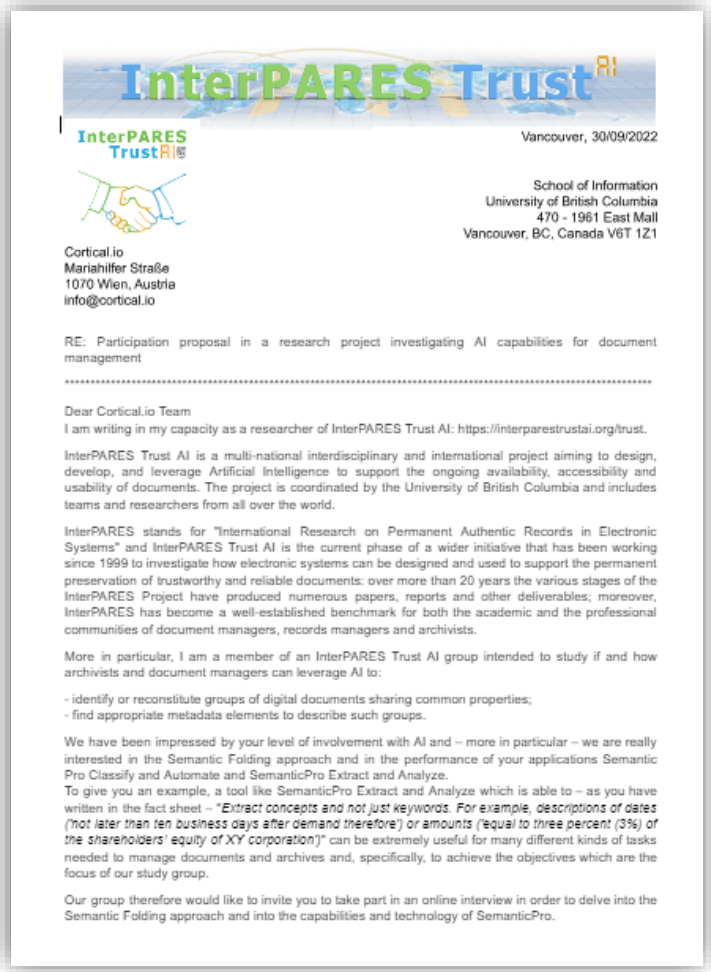
Fase 2. Questionario e interviste

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.**



InterPARES Trust AI

InterPARES Trust AI

Vancouver, 30/09/2022

School of Information
University of British Columbia
470 - 1961 East Mall
Vancouver, BC, Canada V6T 1Z1

Cortical.io
Mariahilfer Straße
1070 Wien, Austria
info@cortical.io

RE: Participation proposal in a research project investigating AI capabilities for document management

Dear Cortical.io Team
I am writing in my capacity as a researcher of InterPARES Trust AI: https://interparestrustai.org/trust.

InterPARES Trust AI is a multi-national interdisciplinary and international project aiming to design, develop, and leverage Artificial Intelligence to support the ongoing availability, accessibility and usability of documents. The project is coordinated by the University of British Columbia and includes teams and researchers from all over the world.

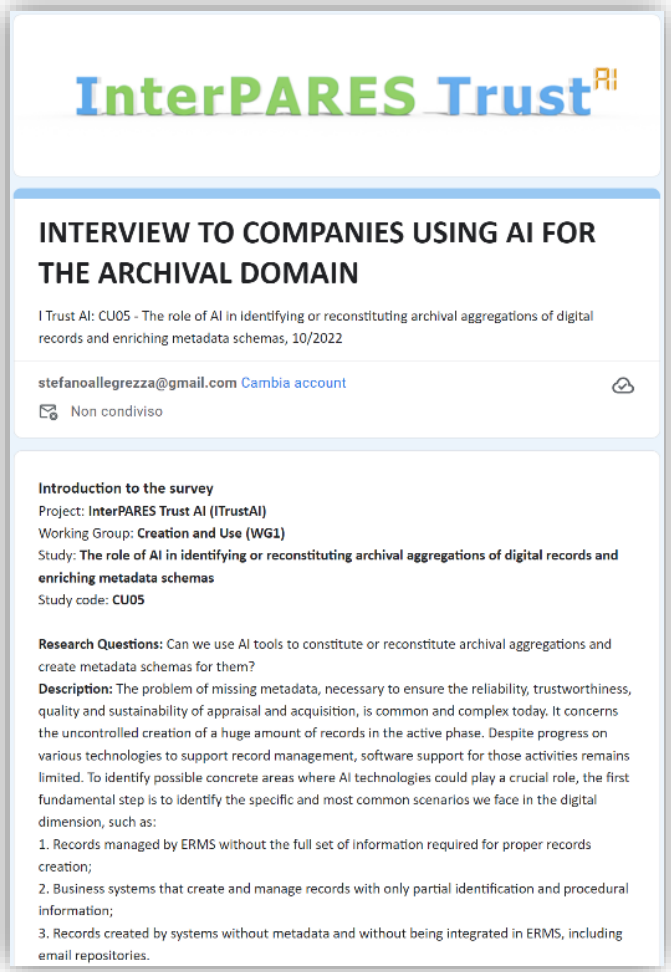
InterPARES stands for "International Research on Permanent Authentic Records in Electronic Systems" and InterPARES Trust AI is the current phase of a wider initiative that has been working since 1999 to investigate how electronic systems can be designed and used to support the permanent preservation of trustworthy and reliable documents: over more than 20 years the various stages of the InterPARES Project have produced numerous papers, reports and other deliverables; moreover, InterPARES has become a well-established benchmark for both the academic and the professional communities of document managers, records managers and archivists.

More in particular, I am a member of an InterPARES Trust AI group intended to study if and how archivists and document managers can leverage AI to:

- identify or reconstitute groups of digital documents sharing common properties;
- find appropriate metadata elements to describe such groups.

We have been impressed by your level of involvement with AI and – more in particular – we are really interested in the Semantic Folding approach and in the performance of your applications Semantic Pro Classify and Automate and SemanticPro Extract and Analyze.
To give you an example, a tool like SemanticPro Extract and Analyze which is able to – as you have written in the fact sheet – *"Extract concepts and not just keywords. For example, descriptions of dates ('not later than ten business days after demand therefore') or amounts ('equal to three percent (3%) of the shareholders' equity of XY corporation')"* can be extremely useful for many different kinds of tasks needed to manage documents and archives and, specifically, to achieve the objectives which are the focus of our study group.

Our group therefore would like to invite you to take part in an online interview in order to delve into the Semantic Folding approach and into the capabilities and technology of SemanticPro.



InterPARES Trust AI

INTERVIEW TO COMPANIES USING AI FOR THE ARCHIVAL DOMAIN

I Trust AI: CU05 - The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas, 10/2022

stefanoallegrezza@gmail.com Cambia account

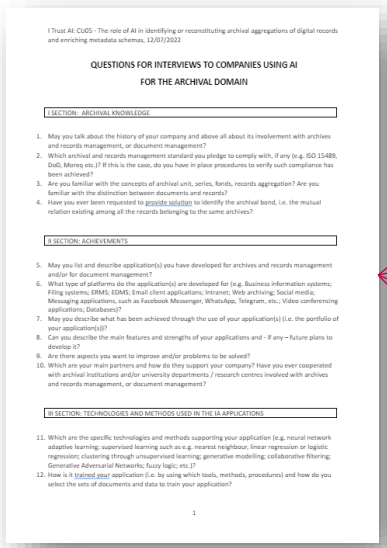
Non condiviso

Introduction to the survey
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**
Study code: **CU05**

Research Questions: Can we use AI tools to constitute or reconstitute archival aggregations and create metadata schemas for them?
Description: 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_7kM4CJptKQStTg/viewform

Fase 2. Questionario e interviste



I Trust AI: CUOS - The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas, 12/07/2022

QUESTIONS FOR INTERVIEWS TO COMPANIES USING AI FOR THE ARCHIVAL DOMAIN

I SECTION: ARCHIVAL KNOWLEDGE

1. May you talk about the history of your company and above all about its involvement with archives and records management, or document management?
2. Which archival and records management standard you pledge to comply with, if any (e.g. ISO 15489, DoD, Moreq etc.)? If this is the case, do you have in place procedures to verify such compliance has been achieved?
3. Are you familiar with the concepts of archival unit, series, fonds, records aggregation? Are you familiar with the distinction between documents and records?
4. Have you ever been requested to provide solution to identify the archival bond, i.e. the mutual relation existing among all the records belonging to the same archives?

II SECTION: ACHIEVEMENTS

5. May you list and describe application(s) you have developed for archives and records management and/or for document management?
6. What type of platforms do the application(s) are developed for (e.g. Business information systems; Filing systems; ERMS; EDMS; Email client applications; Intranet; Web archiving; Social media; Messaging applications, such as Facebook Messenger, WhatsApp, Telegram, etc.; Video conferencing applications; Databases)?
7. May you describe what has been achieved through the use of your application(s) (i.e. the portfolio of your application(s))?
8. Can you describe the main features and strengths of your applications and - if any – future plans to develop it?
9. Are there aspects you want to improve and/or problems to be solved?
10. Which are your main partners and how do they support your company? Have you ever cooperated with archival institutions and/or university departments / research centres involved with archives and records management, or document management?

III SECTION: TECHNOLOGIES AND METHODS USED IN THE IA APPLICATIONS

11. Which are the specific technologies and methods supporting your application (e.g. neural network adaptive learning; supervised learning such as e.g. nearest neighbour, linear regression or logistic regression; clustering through unsupervised learning; generative modelling; collaborative filtering; Generative Adversarial Networks; fuzzy logic; etc.)?
12. How is it trained your application (i.e. by using which tools, methods, procedures) and how do you select the sets of documents and data to train your application?

1

25 questions

I SECTION

II SECTION

III SECTION

IV SECTION

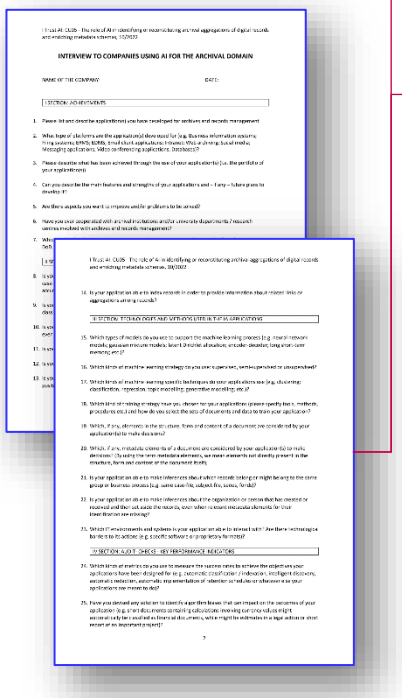
achievements

specific capabilities (for recordkeeping and email systems)

technologies and methods used in the IA applications

audit-checks -- key performance indicators

Fase 2. Questionario e interviste



25 questions

I SECTION

achievements

II SECTION

specific capabilities (for recordkeeping and email systems)

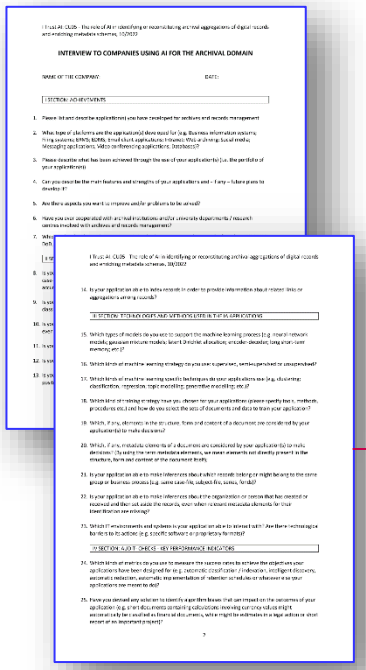
I SECTION: ACHIEVEMENTS

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

III SECTION: TECHNOLOGIES AND METHODS USED IN THE IA APPLICATIONS

15. Which types of models do you use to support the machine learning process (e.g. neural network models; gaussian mixture models; latent Dirichlet allocation; encoder-decoder; long short-term memory; etc.)?
16. Which kinds of machine learning strategy do you use: supervised, semi-supervised or unsupervised?
17. Which kinds of machine learning 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)?

Fase 2. Questionario e interviste



25 questions

I SECTION

achievements

II SECTION

specific capabilities (for recordkeeping and email systems)

III SECTION

technologies and methods used in the IA applications

IV SECTION

audit-checks -- key performance indicators

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

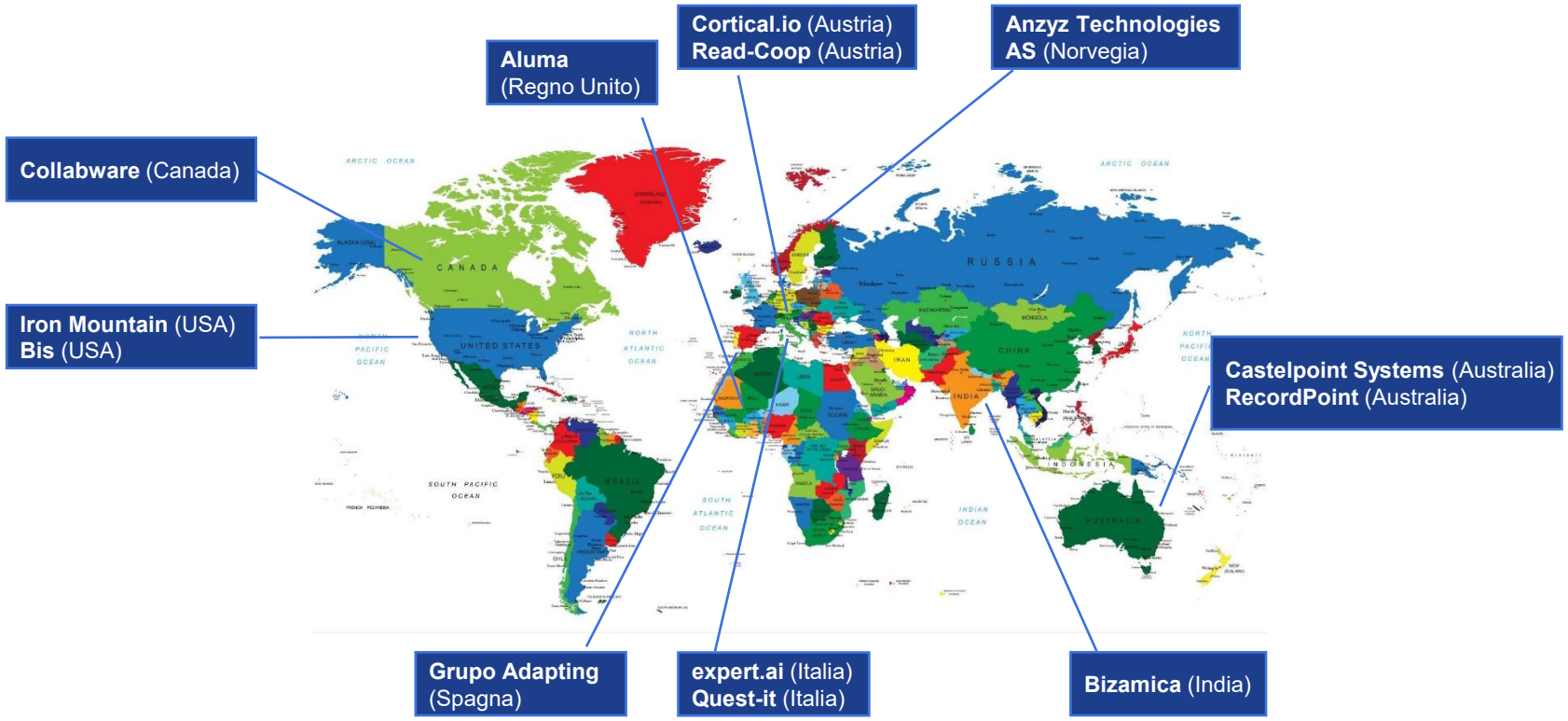
8. Is your application able to file automatically or semi-automatically records in the respective folders, case-files or group they may belong? May it perform this task for both newly created records and accumulated records?
9. Is your application able to classify/file records, folders and groups of 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?

IV SECTION: AUDIT- CHECKS - KEY PERFORMANCE INDICATORS

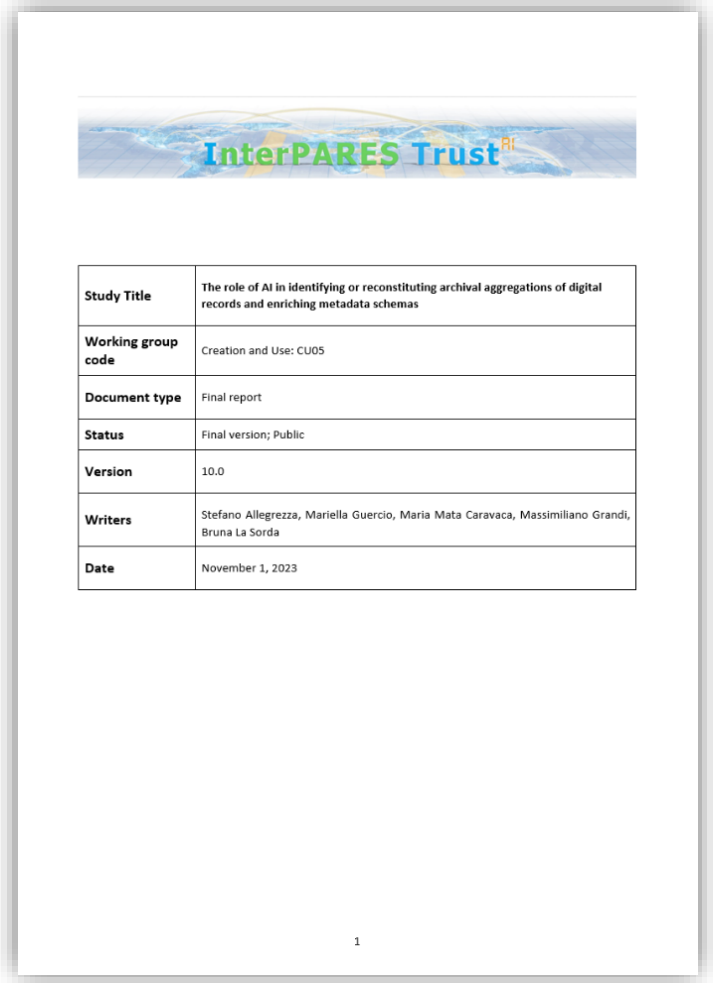
24. Which kinds of metrics do you use to measure the success rates to achieve the objectives your applications have been designed for (e.g. automatic classification / indexation, intelligent discovery, automatic redaction, automatic implementation of retention schedules or whatever else your applications 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)?

Fase 2. Aziende che hanno risposto al sondaggio

Hanno risposto **13 aziende**:



Il rapporto finale



InterPARES Trust AI

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

1

https://interparestrustai.org/assets/public/dissemination/Report-CU05-Survey-of-the-Companies_v121.pdf

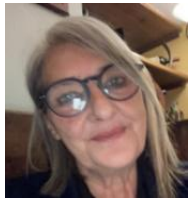
Stefano Allegrezza (co-presidente) (Università di Bologna - Istituto di Ricerca per l'Intelligenza Artificiale Centrata sull'Uomo-ALMA AI)

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)



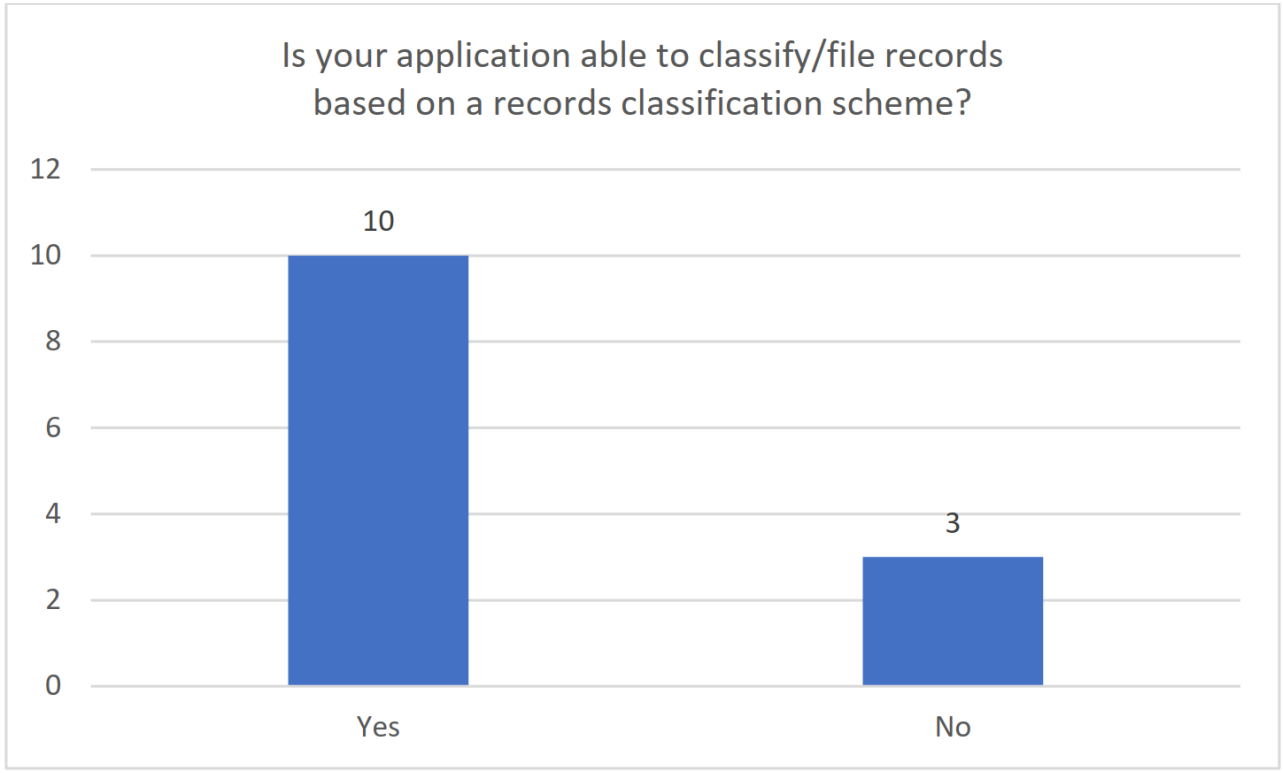
3. Risultati

Alcuni risultati: la classificazione

La **classificazione automatica dei documenti** sulla base di un piano di classificazione è offerta da **quasi tutte le aziende che hanno risposto ai questionari** (10 su 13), comprese quelle non specificamente coinvolte nella gestione di archivi e documenti.

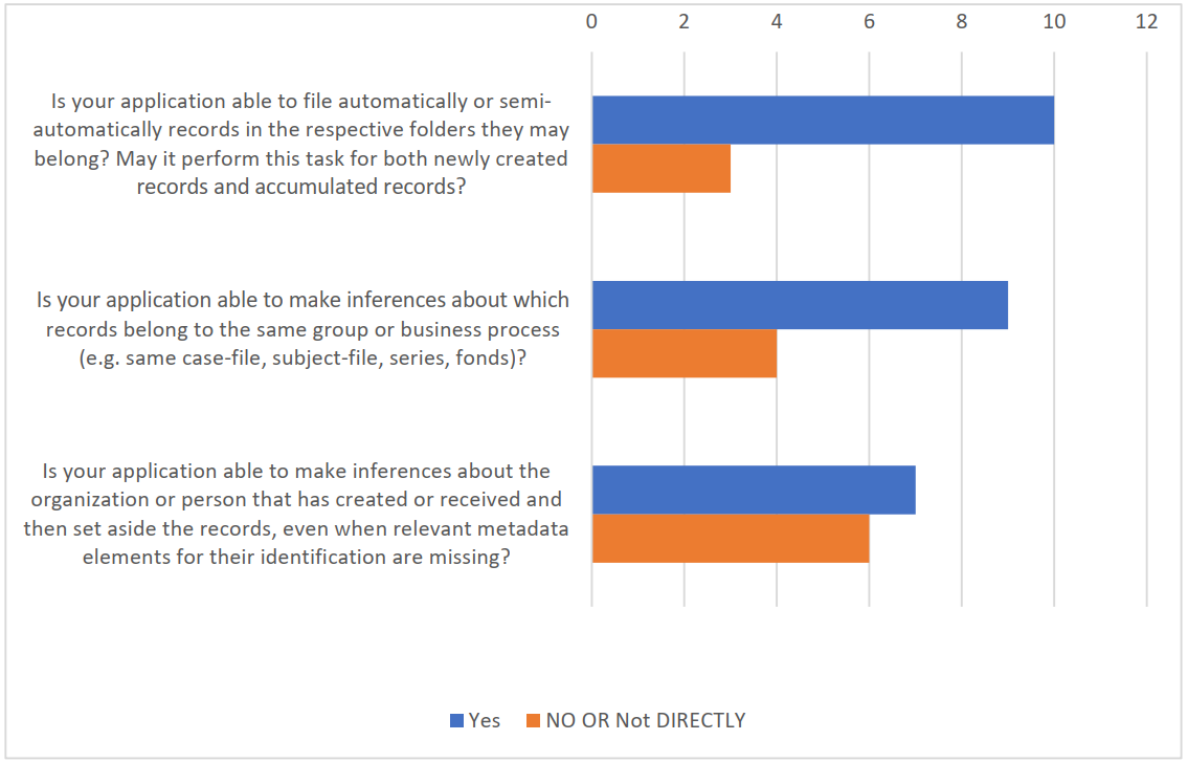
Tra le aziende le cui **applicazioni non prevedono la classificazione automatica**, una si occupa principalmente del riconoscimento del testo manoscritto e le altre due si concentrano sull'indicizzazione e sull'estrazione di elementi di metadati.

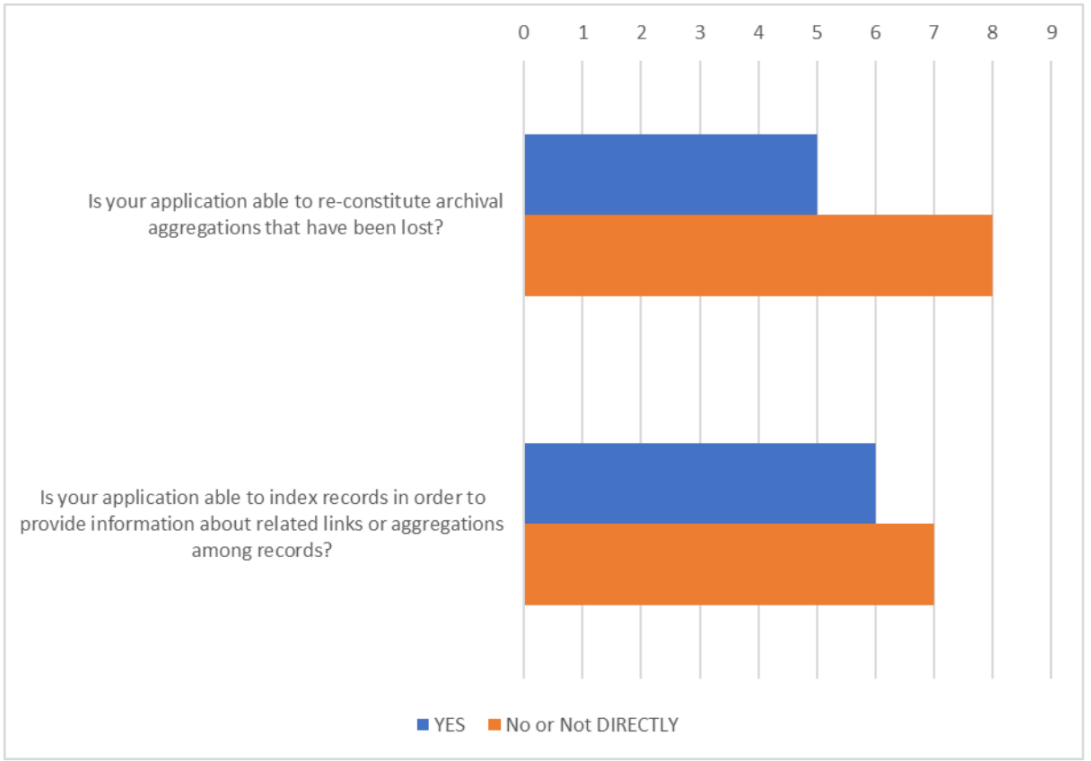
Alcune di queste affermano che le loro applicazioni **potrebbero essere addestrate** per la classificazione dei documenti.



Classificazione

Alcuni risultati: le aggregazioni documentali e il vincolo





Fascicolazione

Ricostituzione del vincolo archivistico

Soluzioni tecnologiche

Techniques and Analysis Models

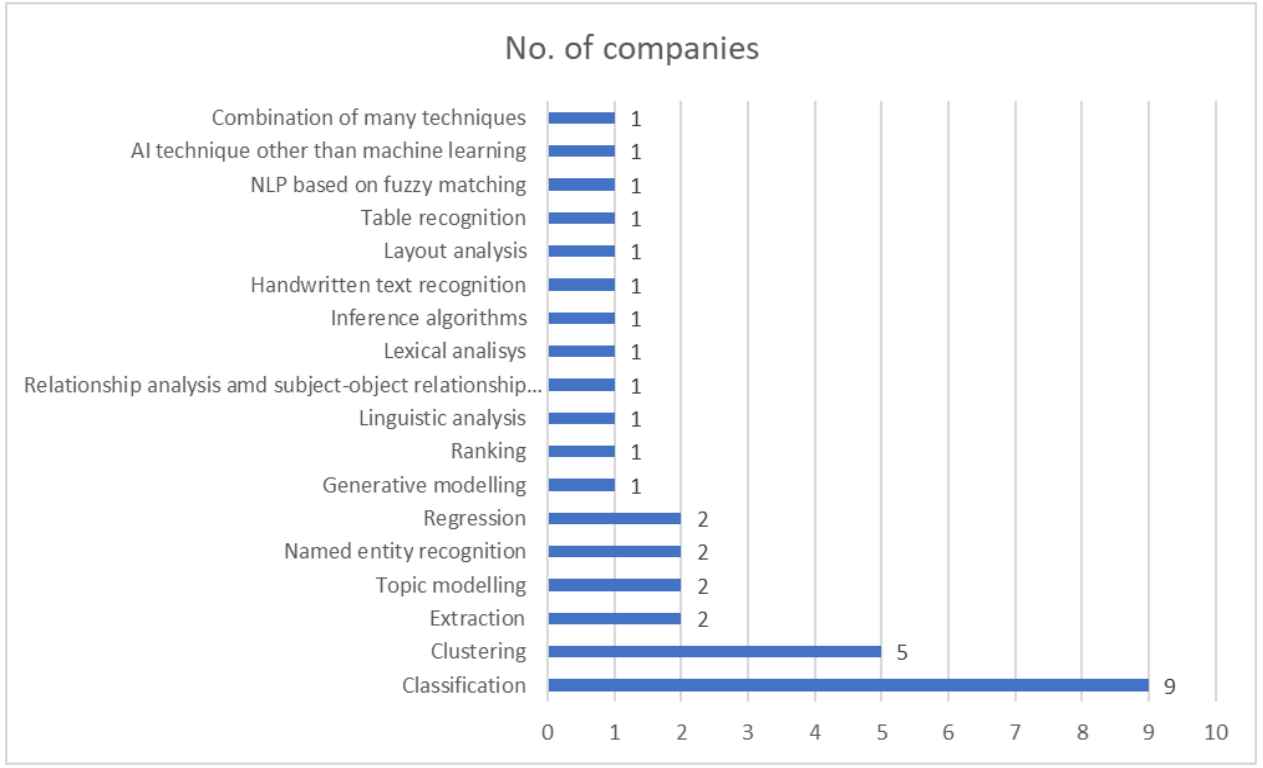
The companies listed a wide range of different **analysis models**: **24 different entries** - the 5 most recurring are:

- **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

Conformità alle norme

Le aziende dichiarano che i loro prodotti sono conformi 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 dei documenti**, 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**,

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.

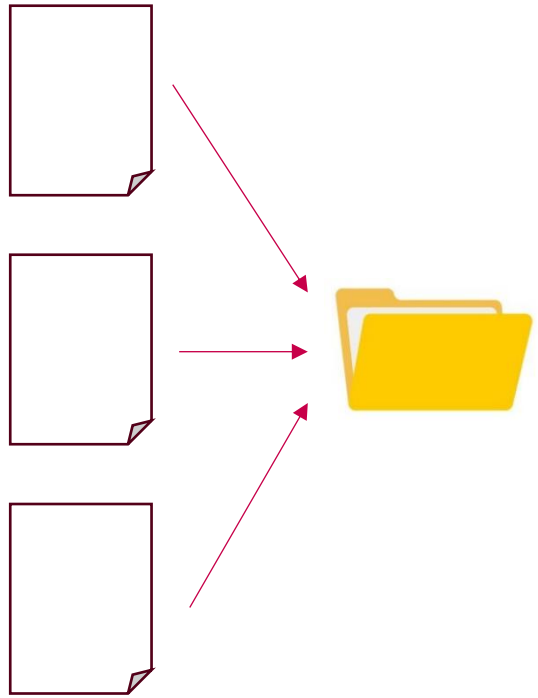
Interessante notare che:

- **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** se si vogliono ottenere risultati affidabili.

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 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 funzionaltà **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 ad un approccio automatico, ma solo **supportata** dalle tecnologie AI attraverso l'**intermediazione degli archivisti**.

Future work

Casi di studio

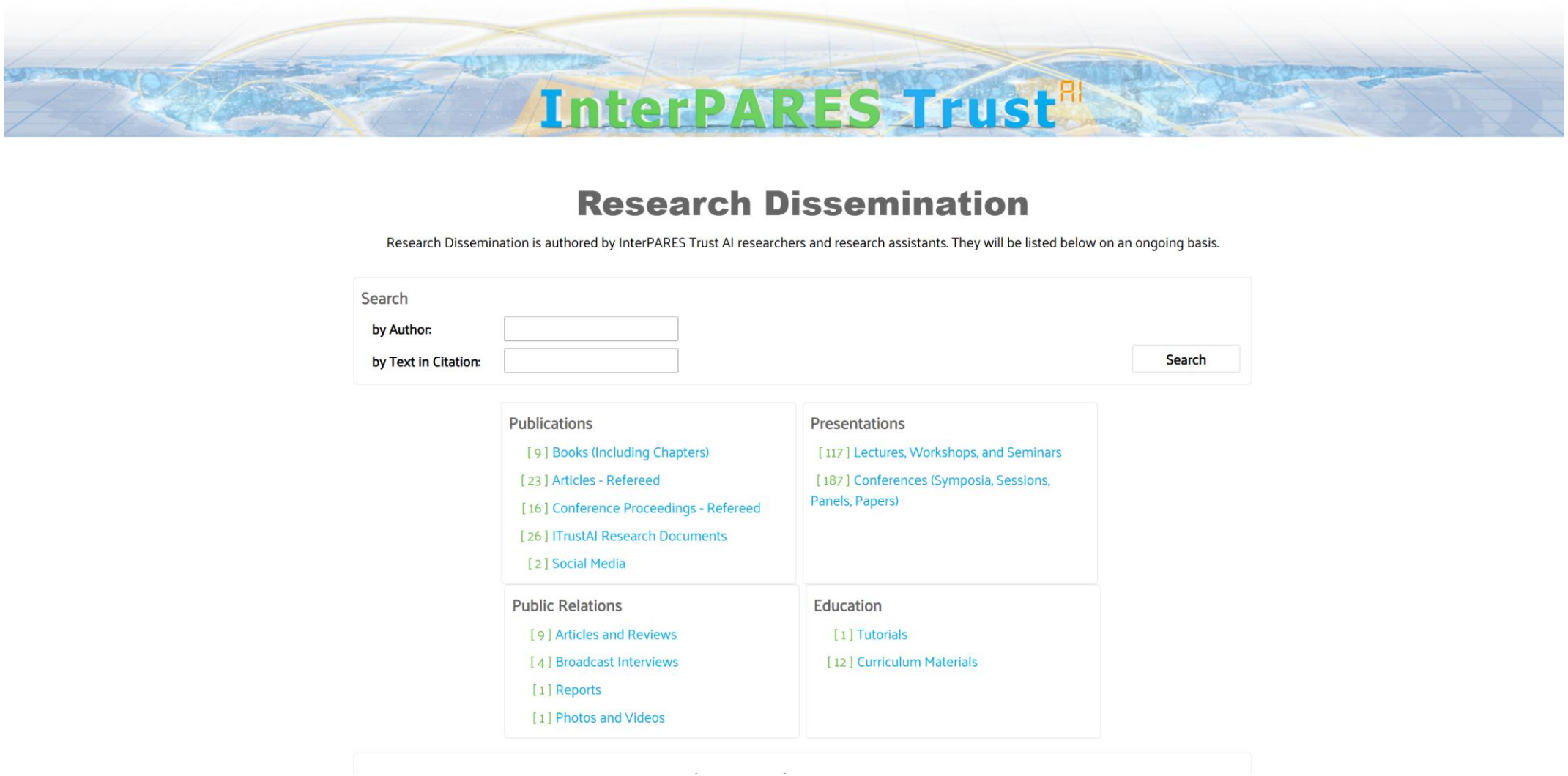
NATO (North Atlantic Treaty Organization)

PAR-ER (Polo archivistico Regione Emilia Romagna)

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?
- Si, 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 le operazioni di **classificazione** e di creazione delle **aggregazioni documentali** potranno (forse) essere svolte integralmente dall'intelligenza artificiale senza intervento umano.

Per saperne di più…



https://interparestrustai.org/trust/research_dissemination

Grazie per l'attenzione!

Stefano Allegrezza
Università di Macerata (ITALIA), Dipartimento di Studi umanistici
stefano.allegrezza@unimc.it