The use of AI in identifying or recreating archival aggregations: a survey on market solutions

Stefano Allegrezza, Mariella Guercio

July, 7 2023

Agenda

1. Introduction: the research question
2. The survey on market solutions: the methodology
3. The survey on market solutions: the questionnaire
4. The survey on market solutions: analysis of answers
5. Conclusions: the survey report

1. Introduction: the research question

The CU05 Study and its main research question

CU05

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

Mariella Guercio (co-chair) (Associazione nazionale archivistica italiana - ANAI)
Stefano Allegrezza (co-chair) (Università di Bologna - Research Institute for Human-Centered Artificial Intelligence-ALMA AI)
Georgia Barloura (European Free Trade Association - EFTA)
Ineke Deserno (North Atlantic Treaty Organization - NATO)
Nicola Di Matteo (Halifax University, Canada)
Georg Gaenser (European Free Trade Association - EFTA)
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)
Leonardo Mineo (Associazione nazionale archivistica italiana - ANAI
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)”)

Can we use **AI tools** to **build** or **recreate archival aggregations** and to **metadata** schemas for them?

Just a couple of examples

In many public administrations and private companies, **documents are neither classified nor aggregated**

In other cases, aggregations of documents are **not well created**, resulting in an uncontrolled number of documents that are **not sorted**, not placed in the correct folder and **difficult to find**.

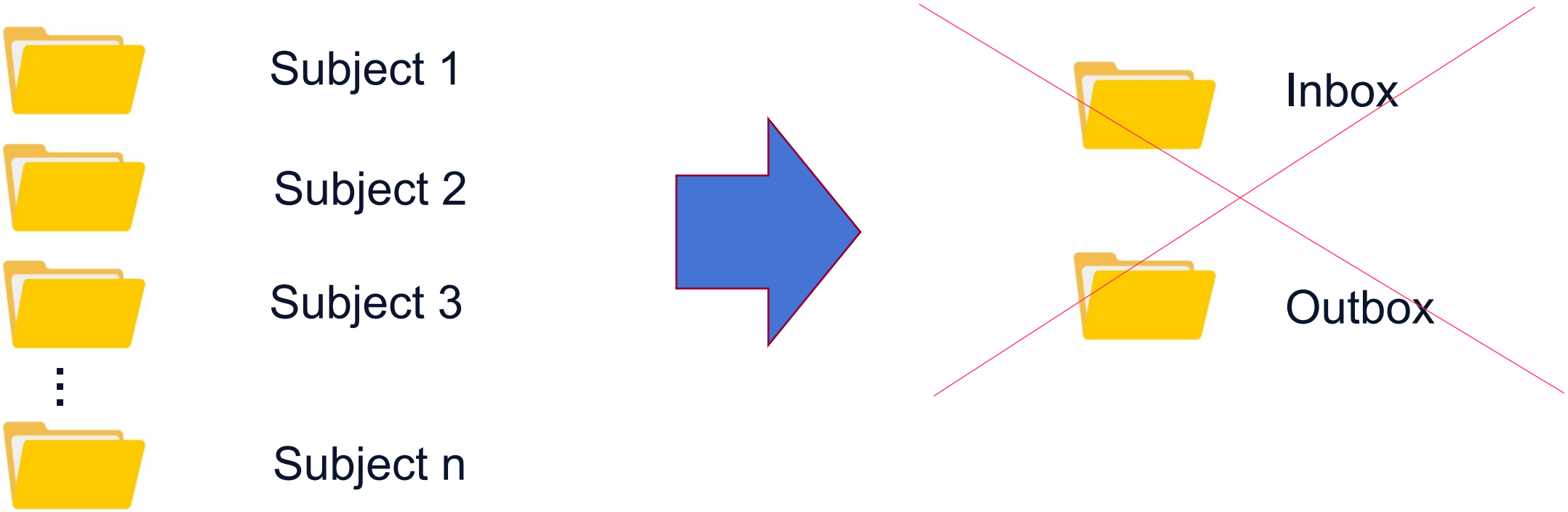
In many cases **metadata -** necessary to ensure the reliability, trustworthiness, quality and sustainability of appraisal and acquisition - **are missing**.

Despite progress on various technologies to support document management, software support for those activities remains limited.

Just a couple of examples

Email management has become **one of the most time-consuming activities** both in the public sector and also in private companies and in personal activities.

Emails are often managed as single records without any bond with other emails and **are not classified or filed in archival aggregations** (folders) nor are connected to and classified in the record management system of the creator.



What AI technologies might be useful under what conditions

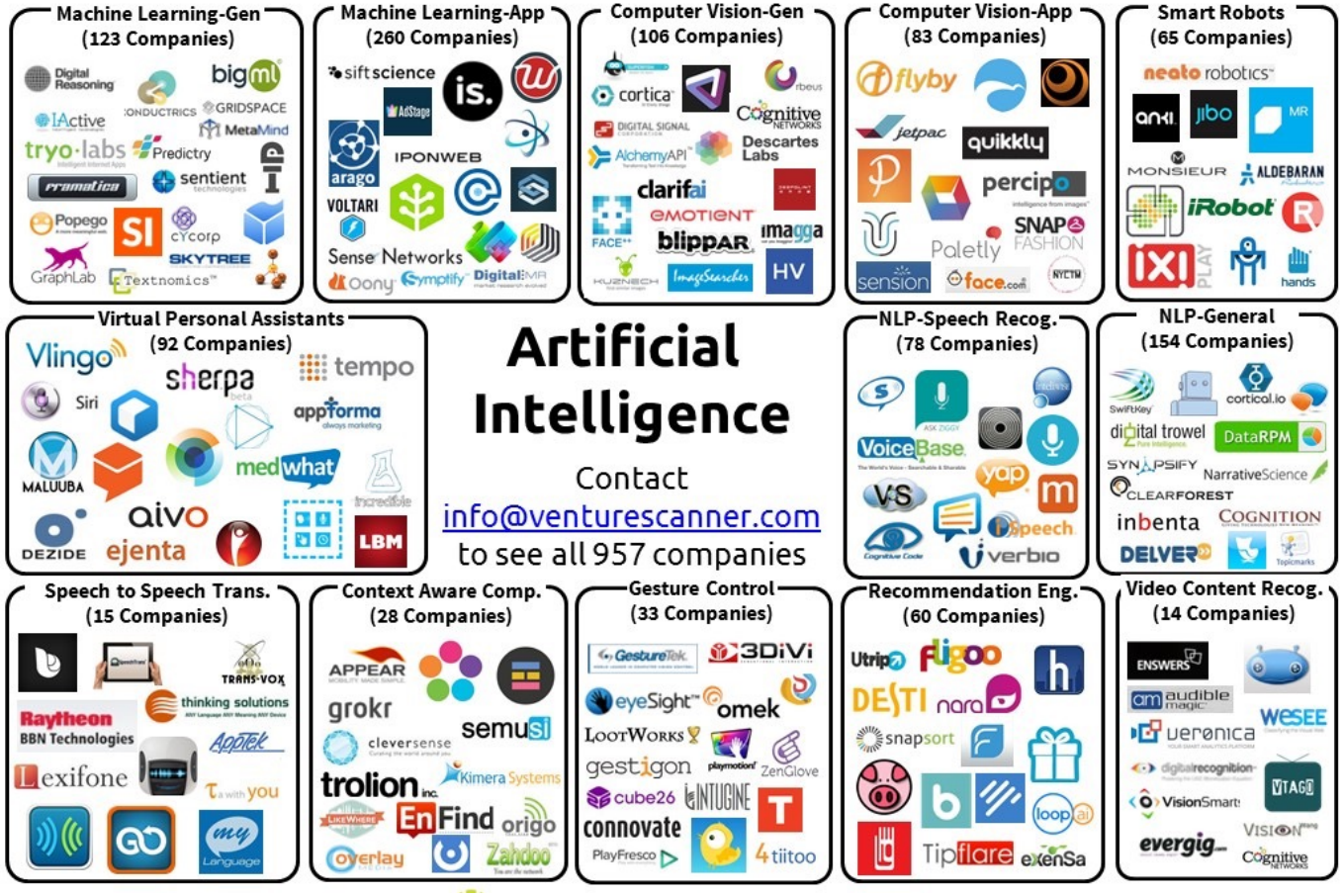
Which **AI technologies** could be useful for this purpose for the automatic or semi-automatic management of emails, for example:

- for automatic classification?
- for aggregating the records?
- for filtering emails?
- for integrating metadata for describing the creation context and use?
- for automatic appraisal and disposal?

The research first step: the analysis of AI software

There are **thousands** of companies that declare they use AI

Hundreds of them declare they use AI-Technologies in the field or ERMS/EDMS



How to verify which archival functions are addressed and their adequacy?

2. The survey on market solutions: the methodology

Phase 1: Identification of AI companies

Identification of an initial group of **300 companies** of interest to the study

Companies that develop IT products and:

- are based on AI-related technologies
- are relevant to the scope of the CU05 study

CU05 The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas

The list is neither exhaustive nor definitive, but a **starting point**

Tools for building the list:

- **direct Internet searches** using keywords and text strings;
- **resources and knowledge made available by professionals**
  (Alan Pelz-Sharpe, Andrew Warland, James Lappin, Jenny Bunn and Paul Young)

Phase 1: Elements for a preliminary evaluation

The group was later limited to **100 companies**

The features of their **AI software** were first analyzed according to the **information available on their websites** with reference to:

- **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 it is not openly stated but can only be guessed from the contents of the website).

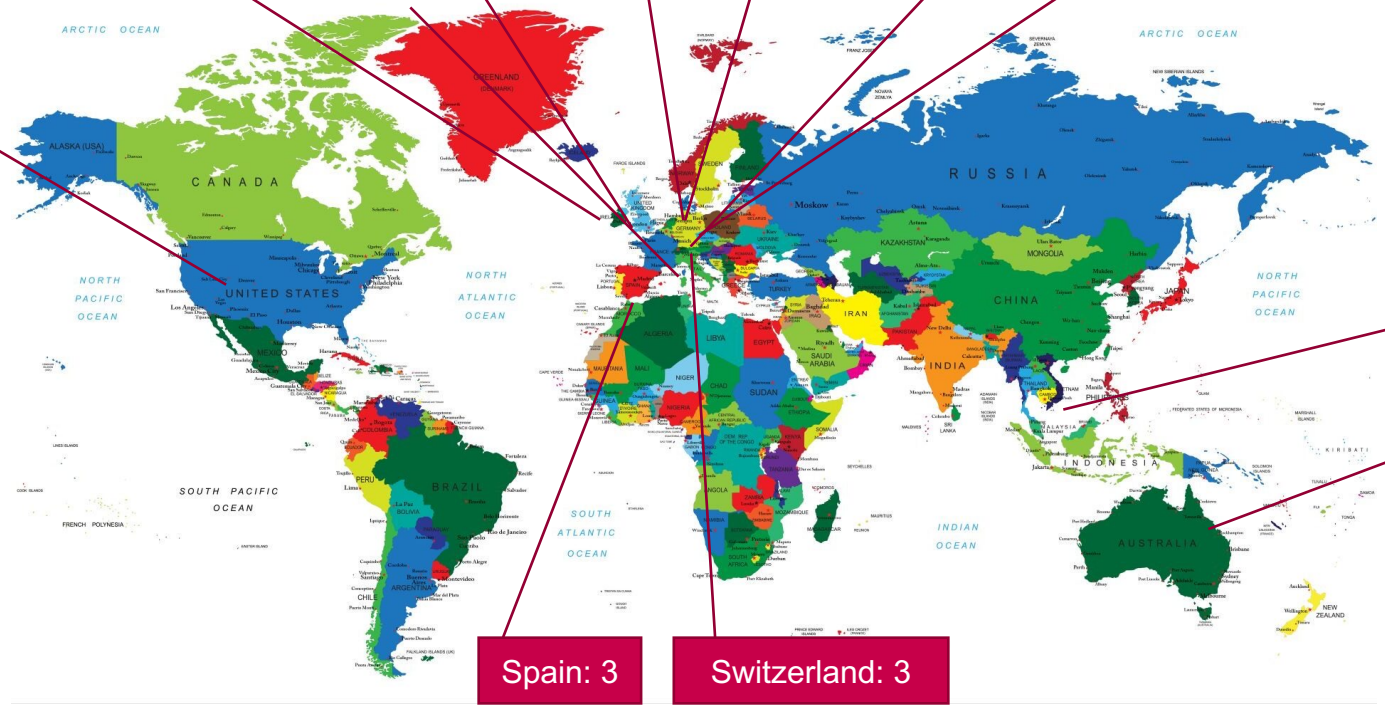
LIST OF 100 COMPANIES THAT ARE OR MIGHT BE RELEVANT TO CU05

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 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

The **notes** describe the most interesting products developed by the company, detail statements made by the company itself about their commitment to objectives relevant to CU05 and contain any other information explaining why what the company does is or may be of interest to CU05

Phase 1: 100 companies – geographical distribution

Geographic location always refers to the main and/or **original location of the company** (often a company has multiple locations):



Germany: 5

Austria: 3

Netherlands 4

UK: 10

Ireland: 2

Belgium: 2

France: 2

USA: 46

Singapore: 2

Australia: 4

Spain: 3

Switzerland: 3

Brazil: 1
Bulgaria: 1
Canada: 1
Colombia: 1
Cyprus: 1
Czech Republic: 1
Finland: 1
Lithuania: 1
New Zealand: 1
Portugal: 1

Phase 1: Identification of AI companies

Since it was not possible to interview all **100 companies**, from the initial list we selected a list of **28 companies** on the basis of:

- their **portfolio**
- their direct involvement in the **record field**
- their **compliance** with regulatory frameworks and standards relevant in the domain
- the general **reputation** of the company.

It is best to avoid talking to sales representatives and instead contact **information management personnel**, **software engineers**, and **archivists** (if any).

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	Sienna, Italy	www.quest-it.com
7	Grupo Adapting	Valencia, Spain	https://www.adapting.com/en/
8	Hyland	Westlake, Ohio, USA	https://www.hyland.com/en
9	Stratagem	Aurora, Colorado, USA	www.stratagemgroup.com
10	Aluma	Cambridge, UK and New York, USA	https://aluma.io/
11	Collabware	Washington, DC, USA	collabware.com
12	Ephesoft	Irvine, California, USA	https://ephesoft.com/
13	Read-Coop	Innsbruck, Austria	https://readcoop.eu/transkribus/
14	Recordpoint	Sydney, Australia	www.recordpoint.com
15	Prism Software	California, USA	https://prismsoftware.com/
16	ExpertSystem	Modena, Italy	https://www.expert.ai/
17	GRMdocument management	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	Gleematic AI	Singapore	https://gleematic.com/
25	SBK Business Solutions	São Bernardo do Campo, São Paulo, Brazil	www.sbkbs.com.br
26	Datacentrix	Johannesburg, South Africa	www.datacentrix.co.za

\+ Anzyz (Norway)
DXC (Italy)

Phase 2. Questionnaire and interviews

In order to gather more precise information, we prepared a very detailed **questionnaire** aimed at collecting systematically the information for an adequate assessment of the applications

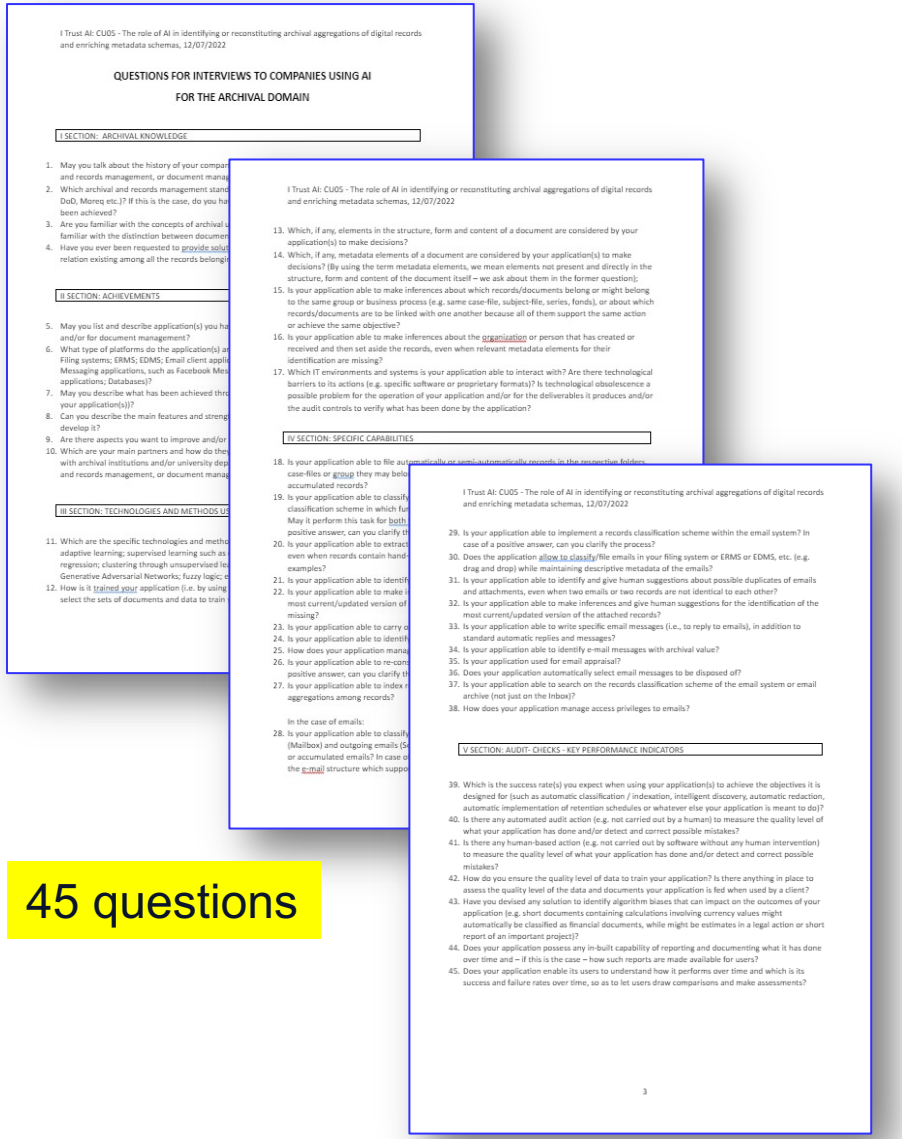
We sent to the **28 companies** an official **invitation letter** (in English, in Spanish or Portuguese, according to the preferred language of the company) to take part in the survey

The questionnaire was explained orally during a **preliminary meeting** with information management staff and software engineers.

Subsequently, the companies filled out the questionnaire available on **Google Forms**

https://docs.google.com/forms/d/e/1FAIpQLSc8US3a89JbVjhfdma2EqYm1Xo_LVqP3qh_7kM4CJptKQStTg/viewform

Phase 2. Questionnaire and interviews



I Trust AI: CU05 - 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

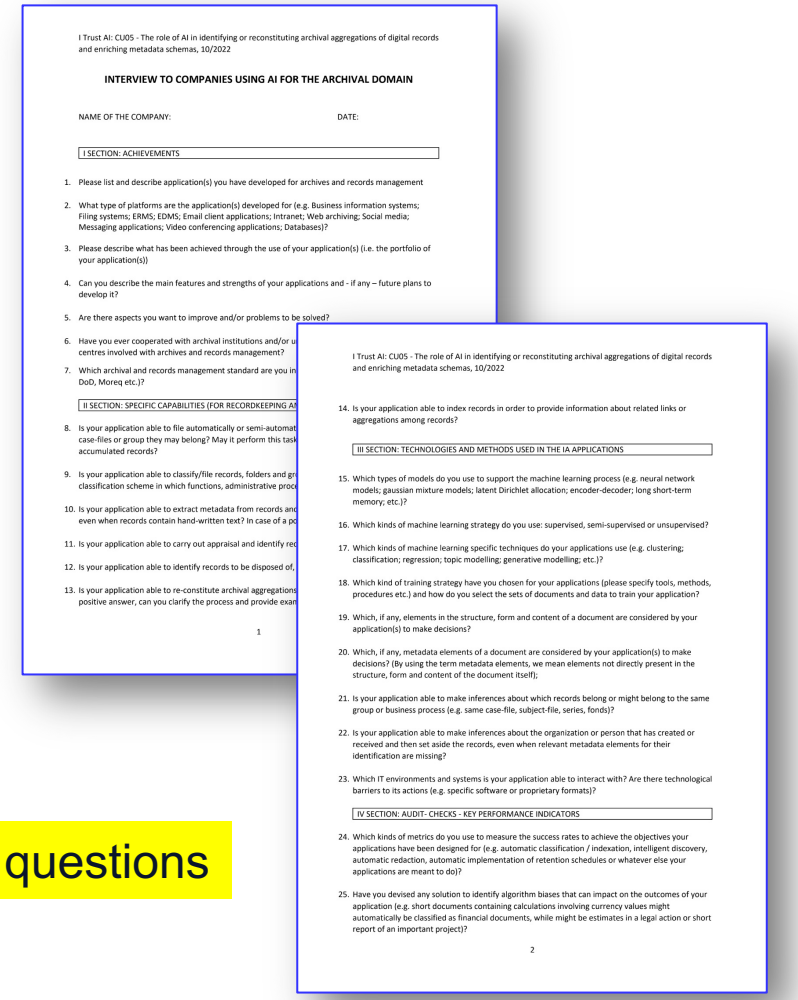
II SECTION: ACHIEVEMENTS

III SECTION: TECHNOLOGIES AND METHODS US

IV SECTION: SPECIFIC CAPABILITIES

V SECTION: AUDIT- CHECKS - KEY PERFORMANCE INDICATORS

45 questions



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

INTERVIEW TO COMPANIES USING AI FOR THE ARCHIVAL DOMAIN

NAME OF THE COMPANY: DATE:

I SECTION: ACHIEVEMENTS

II SECTION: SPECIFIC CAPABILITIES (FOR RECORDKEEPING AP

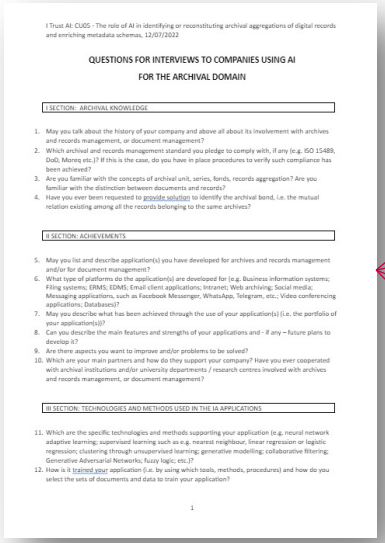
III SECTION: TECHNOLOGIES AND METHODS USED IN THE IA APPLICATIONS

IV SECTION: AUDIT- CHECKS - KEY PERFORMANCE INDICATORS

25 questions

3. The survey on market solutions: the questionnaire

The questionnaire: the sections



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 → 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

The questionnaire: the questions

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; Messa
3. Please your a
4. Can yo develo
5. Are th
6. Have y centre
7. Which DoD, M

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)?

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)?

Companies that accepted the survey

13 companies replied:

Collabware (Canada)

Iron Mountain (USA)
Bis (USA)

Aluma (UK)

Cortical (Austria)
Read-Coop (Austria)

Anzyz Technologies AS (Norway)

Castelpoint Systems (Australia)
Recordpoint (Australia

Groupo Adapting (Spain)

Expert.ai (Italia)
Quest-it (Italia)

Bizamica (India)

4. The survey on existing software: analysis of answers

The portfolio of the companies

All the market players interviewed have developed solutions based on AI technologies for indexing and/or classifying structured, semi-structured and unstructured data/records based on **automatic learning techniques** and **automatic data extraction**.

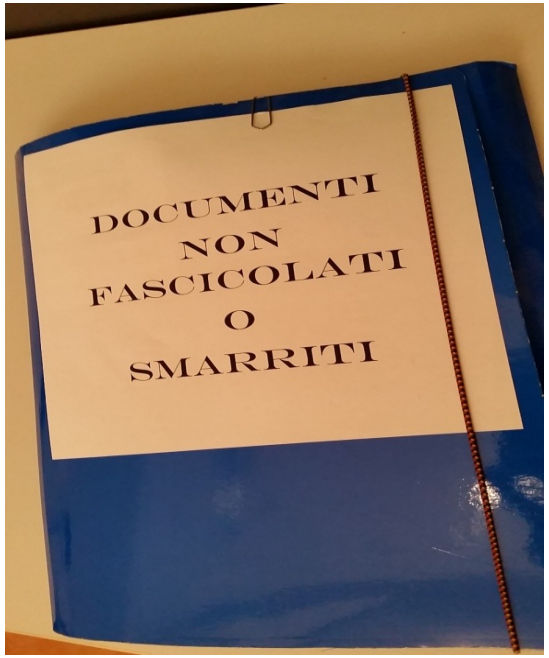
The amount of specific services listed is huge, detailed and diversified:

- not necessarily these peculiarities testify **approaches really different;**
- could the **variety of creative solutions** be the consequence of the complex tasks required for **respecting the peculiarities of the archival requirements?** or
- does it reflect the intrinsic nature of dynamic technologies still dominated by an ongoing process of evolution and transformation?

Survey outcome from the archival perspective (1/4)

The majority of the market players interviewed have proved:

- to be able to **understand the complexity and the relevance of archival environment and functions**
- to be aware of the **uniqueness of the original metadata acquired in the creator's current activities**, both if the issue concerns the records' automatic classification or in case of the creation of archival aggregations.

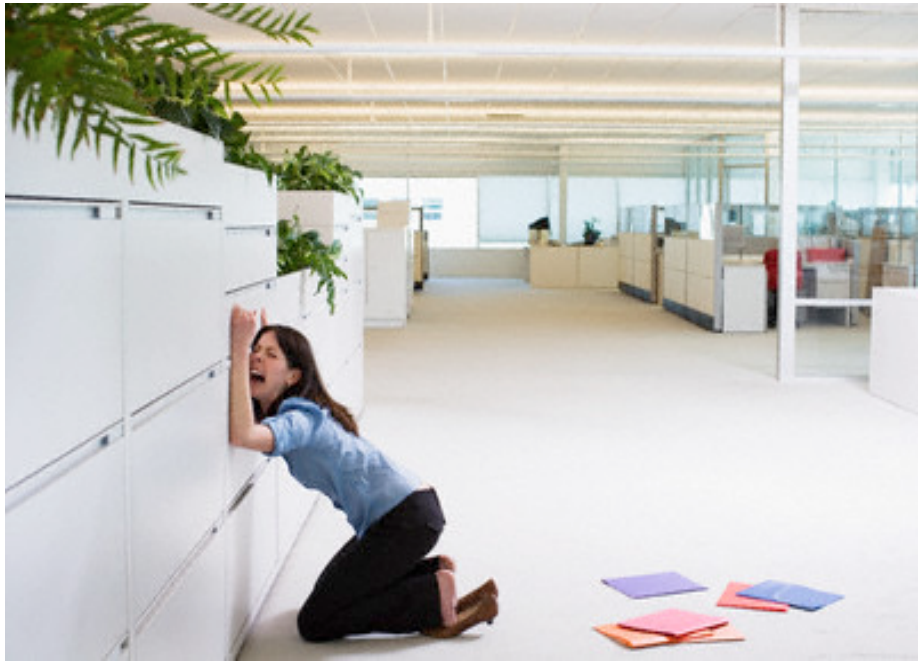


not filed or lost records

Survey outcome from the archival perspective (2/4)

- The role of any **metadata fields found or inferred** is always at the center of any reply.
- 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 *archival* 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 is considered not replaceable** for providing consistent results.



Survey outcome from the archival perspective (3/4)

In terms of **records aggregation** or **re-aggregation**, the promises for automatization are not very encouraging, as this possibility is confirmed to be **limited to very specific cases** such as

- **defining records types**, when the **users' specifications are already in place**, or
- **establishing functional relations among records** when the **original structure** of the content source already provides **basic *intelligent* information**.

The automatic or semi-automatic aggregation based on the document content is **only suggested** and is usually **supported by user validation**, of **human-in-the-loop workflow** or **rules availaible at the creation**

- in more cases even these **limited capacities are not already developed** but **in the process** of being developed.

Survey outcome from the archival perspective (4/4)

Even the **provenance information seems not easily recognizable** 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.

Also the **reconstitution of the archival bond** – when lost or not explicitly defined – is recognized as a complex activity, 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**

Remarks from the archival perspective (1/2)

The survey shows for all respondents a **cautious approach** when questions concern records and contextual relationships with the archive.

The reasons could depend:

- on the **strict parameters** we have adopted for selecting the market companies, but also
- on the **degree of interactions and explanations exchanged between the researchers and the companies** involved in the review during the questionnaire submission.

A real concern from the providers or intimidation from archivists?

Remarks from the archival perspective (2/2)

- In any case, the experience matured testifies that the complexity of archival functions cannot be easily reduced and removed by an automatic approach, but only supported by the AI technologies through the **intermediation by users and professionals**.
- The **terminology is a crucial challenge**.
- As a consequence, when interacting with market players involved in the implementation of AI platforms in the records and archival domains, the archival community must pay a lot of attention
  - to clarify their concepts behind general terms such **aggregation** and **classification** and
  - to correctly interpret AI expression such, "**Intelligent Document Processing**" which, usually, has nothing to do with document, with its processing and, at the end, with ***archival intelligence*.**

5. Conclusions: the survey report

The survey report (draft)

The results of the questionnaire will be included in a **final report** that is currently being drafted and is expected to be completed by the end of august.

TABLE OF CONTENTS

Thank you!

Any comments are welcome!

Stefano Allegrezza and Mariella Guercio