Title	**The role of AI for records management: findings from case studies**
Working group code	CU05
Study title	
Status	Final
Version	
Writers	Stefano Allegrezza, Gabriele Bezzi, Maria Mata Caravaca, Massimiliano Grandi, Mariella Guercio, Bruna La Sorda, Francesca Magnoni, Marianna Tascone
Date	February 2026

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

The role of AI for records management: findings from case studies

Summary

1. Introduction

The study CU05 “The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas” has been developed in two steps. The first one was a survey of the marketplace with specific reference to those companies whose products and initiatives have explicit or implicit relevance to documents and records management. The outputs are described in chapter 2 of this report, which summarizes the survey methods and steps, the results and the analysis of the potentialities of AI models and their limits. The complexity and the dynamic nature of the environment have required a second phase of the research which has been carried out by the team listed at chapter 3 and has been described in this report dedicated (chapter 4) to the analysis of the role of AI for records management supported by concrete case studies. More specifically, the research conducted in this second phase examines and documents AI models quality, efficiency and limits in relation to the functions applied for defining or reconstituting the archival bond. The methodology applied has been further specified in Chapter 5. Chapter 6 dedicated to case studies also includes preliminary considerations (6.1) on the use and the integration of the paradata framework developed within the InterPARES TRUST AI research to document the AI projects and used for describing in detail the NATO case study (6.2). The case study carried out by Regione Emilia Romagna (6.3) has investigated the possibility of automating the assignment of classification elements (classes) to administrative records. A third case concerns the archival documentation of CISU (Centro Italiano Studi Ufologici). The findings (chapter 7) and the conclusions (chapter 8) are related to the overall CU05 study project. The findings present a concise analysis of the functions that AI can support in the field of records and archival management, along with the essential conditions of use that must be observed to maintain quality control over the processes and procedures ensuring the integrity, accuracy, and reliability of documentary and information systems across both public and private sectors. The conclusions (chapter 8) are organized in four sections: specific methodological remarks on classification, records aggregation and indexation (8.1), the main achievements from both market survey and case studies (8.2), the description of the information framework used in this report for documenting AI project and the role of recordkeeping systems to maintain it overtime (8.3) and final recommendations (8.4). A list of main references (chapter 9), a brief indication of possible next steps (chapter 10) and of dissemination initiatives (chapter 11) conclude the report. Annex 12.1 describes in detail the framework adopted for documenting the case studies and includes all the cases (12.2. NATO, 12.3. Regione Emilia Romagna, 12.4. Centro Italiano Studi Ufologici – CISU).

2. Summary and Results of Phase 1

The first phase of the InterPARES Trust AI – Creation and Use (CU05) study entitled “The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas” was conducted to assess the actual and potential contribution of Artificial Intelligence (AI) technologies to core archival and records management functions, with particular emphasis on the identification, creation, retrieval, and enrichment of archival aggregations and recordkeeping metadata in digital environments. The study was motivated by widely recognized and increasingly critical problems affecting both public administrations and private organizations: the uncontrolled growth of digital records that are frequently unclassified, unaggregated, and de-contextualized. In many organizations, records are created and archived without systematic application of classification schemes, file plans, or metadata standards. As a result, records often became difficult to retrieve, appraise, manage, and preserve over time. Study CU05 addressed the question of whether artificial intelligence technologies are more advanced than traditional document management tools and can support or partially automate the creation and reconstruction of archival relationships and enriching metadata in a manner consistent with archival principles.

From a conceptual perspective, the study was firmly grounded in archival theory. It explicitly recognized the importance of the archival bond, original order, and contextual relationships as foundational elements for ensuring the authenticity, reliability, and trustworthiness of records. The research did not assume that AI could or should replace archival expertise; rather, it investigated whether AI could support archivists and records managers in dealing with scale, complexity, and loss of control typical of contemporary digital environments. The study therefore aimed not only at identifying existing technological capabilities, but also at highlighting limitations, risks, and requirements for future AI development aligned with archival needs.

Methodologically, the CU05 study adopted a *structured and multi-phase approach*. An extensive market scan was carried out between February and August 2022, during which approximately 300 companies active in the AI and intelligent document processing sectors were analyzed. From this initial pool, 28 companies were selected based on explicit or implicit relevance to document and records management, declared interest in archival contexts, and perceived maturity of their solutions. Thirteen companies ultimately agreed to participate in the study. These companies represented a diverse range of geographical regions, organizational sizes, and technological approaches, including both long-established records management providers and specialized AI firms. Data collection relied on a detailed questionnaire developed by the CU05 team, complemented by online interviews with company representatives. The questionnaire addressed four main areas: company achievements and product characteristics; capabilities relevant to record management and archives; underlying AI technologies and training strategies; and performance measurement, auditability, and bias management. This approach enabled the research team to collect systematic, comparable information while also allowing companies to explain their solutions in depth and contextualize their claims.

One of the most significant results of the study concerned automatic classification capabilities. Most surveyed companies demonstrated that their AI-based platforms were able to classify records automatically or semi-automatically, using combinations of metadata analysis, content-based techniques, and predefined rules. Classification was often configurable to reflect organizational taxonomies or records classification schemes, and in several cases could be adapted by users through training or rule definition. This finding represented a major achievement, as classification is a foundational activity in records management and a prerequisite for many downstream processes, including aggregation, appraisal, and retention.

Closely related to classification, the study found substantial progress in the automated aggregation of records. Most of the surveyed solutions were able to group records into folders, case files, or other logical aggregations based on document type, shared metadata, content similarity, or contextual indicators. Several platforms could infer relationships among records belonging to the same business process or case, particularly when identifiers such as case numbers, transaction references, or shared actors were present. While these aggregation mechanisms were not always explicitly articulated in archival terms, they nevertheless demonstrated a concrete ability to support the reconstruction or approximation of archival groupings in digital systems.

The **reconstitution of the archival bond** was identified as one of the most challenging aspects of the study. Only a limited number of companies declared that their solutions were able to actively support the reconstruction of lost or disrupted aggregations. Where such capabilities existed, they were primarily based on clustering techniques, similarity detection, and inference from contextual or content-based elements. The study concluded that full and reliable reconstitution of archival bonds remained difficult, especially when contextual metadata had been lost entirely. Nevertheless, the ability of some AI tools to propose plausible aggregations or relationships was recognized as an important and promising result, particularly in scenarios involving legacy systems or poorly managed digital repositories.

Metadata extraction and enrichment emerged as one of the strongest and most mature areas identified by the CU05 study. All surveyed companies reported advanced capabilities for extracting metadata from records, including descriptive, administrative, and technical elements. Techniques such as natural language processing, named entity recognition, semantic analysis, and optical character recognition were widely employed to derive metadata from textual content, including scanned and, in some cases,

handwritten documents. This widespread capacity for metadata enrichment was considered a major achievement, as it directly supports improvement, retrieval, governance, and long-term preservation of digital records.

Regarding **appraisal and retention**, the study revealed a heterogeneous landscape. Approximately half of the surveyed companies indicated that their platforms supported appraisal-related activities, often indirectly through classification outcomes, scoring mechanisms, or business rules. Several solutions were able to apply retention schedules automatically once records had been classified, thereby supporting compliant and auditable disposition processes. These capabilities were particularly relevant in regulated environments and demonstrated that AI could effectively support lifecycle management when embedded within clear governance frameworks. However, appraisal as a concept rooted in archival judgment was not always fully understood or explicitly implemented by vendors.

The study also documented the wide range of **AI technologies and training strategies** used by the participating companies. Supervised learning was the most common approach, often complemented by unsupervised, semi-supervised, self-supervised, or rule-based techniques. Companies employed a broad spectrum of models, including neural networks, support vector machines, decision trees, clustering algorithms, and hybrid approaches combining statistical and symbolic methods. In many cases, users or customers played an active role in training and refining models, underscoring the importance of domain knowledge and human-in-the-loop approaches in achieving reliable results.

Another important result concerned compliance with archival, records management, and information governance standards. Several companies explicitly stated that their products supported or were designed in alignment with standards such as ISO 15489, ISO 16175, ISO 23081, ISO 30301, MoReq2010, DoD 5015.02, and GDPR. This demonstrated a growing awareness within the AI industry of regulatory and professional requirements and suggested a positive convergence between technological innovation and archival frameworks.

In conclusion, the CU05 survey indicated that AI technologies have already achieved meaningful and practical results in supporting records management and archival functions, according to the information provided by participating companies. The most significant achievements included automated classification, large-scale metadata extraction and enrichment, support for record aggregation, enhanced search and discovery, and partial automation of retention and compliance processes. At the same time, this first phase of the study highlighted persistent limitations, particularly in the reliable reconstitution of archival bonds and in the explicit incorporation of archival concepts into AI system design. Overall, the study concluded that AI represented a powerful enabling technology rather than a substitute for archival expertise. When appropriately designed, configured, and governed, AI tools were shown to support archivists and records managers in addressing the scale, complexity, and fragility of digital records. The CU05 study emphasized the need for continued collaboration between archival professionals and AI developers to ensure that future solutions were not only technologically advanced, but also conceptually sound, ethically responsible, and aligned with the public interest.

3. Study Team

The study was carried out by the following researchers:

- Stefano Allegrezza (co-chair) (Università di Macerata, Italy)
- Mariella Guercio (co-chair) (Associazione nazionale archivistica italiana - ANAI)
- Gabriele Bezzi (Associazione nazionale archivistica italiana - ANAI)
- Lluís-Esteve Casellas Serra (Ajuntament de Girona, Spain)
- 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 the European Union - European University Institute)
- Gianni Penzo Doria (Università di Udine, Italy)
- Marianna Tascone (Regione Emilia-Romagna)

4. Goals

The second and last phase of CU05 study dedicated to the role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas has focused on specific goals based on the results and considerations obtained in the first part of the research project. More specifically, the researchers have concentrated their efforts on the following activities:

- verifying through concrete case studies the consistency of the archival methodological approach assumed in the first phase of the study,
- analyzing the quality and the efficiency of AI models, when concretely applied for automatic or semi-automatic classification, records aggregation and archival indexing,
- defining conditions and limits for training AI models,
- describing and documenting AI projects by using (but also, if necessary, integrating) the Paradata Framework developed by RP04 Study on “Preserviang AI Techniques as Paradata”.

5. Methodology

The methodological assumptions of CU05 study are consistent with the basic principles of the archival discipline and its main goal: the reliable creation of records and their maintenance over time in their integrity, authenticity, comprehensibility, and reusability independently from the technological and organizational contexts AI models included.

Based on these assumptions, the market analysis carried on in the first part of the research was developed with the aim of creating “an investment guide, highlighting the need for caution in selecting applications and using them in compliance with the principles of reliability and authenticity of documentary production, both of which are essential for assets with legal significance (documents, archival aggregations)[1].” The methodology used allowed us to identify companies whose business plans were consistent to the archival principles and to select partners for implementing, on a trial basis, the concrete application of AI in the records management systems, with specific attention for automatic classification and creation of functional archival aggregations.

The work in the second phase specifically focused on the use of Document AI and on its specific characteristics and functions and was based on analyzing concrete case studies. Document AI refers to the field of AI that manages functions relevant to document and archive management, particularly activities dedicated to processing, understanding, and managing documents using machine learning algorithms, but also the natural language processing (NLP) tools. This, at least on paper, yields advantages in terms of:

- increased efficiency,
- reduced human errors,
- improved search and, therefore, the ability to access information,
- automatic document organization,

[1] *Intelligenza artificiale e archivi digitali. Interazioni e governance,* M. P. Giovannini ed. (2025), Padova: Associazione Siav Academy, 2025, p. 133, <https://www.associazionesiavacademy.it/it/intelligenza-artificiale-gestione-documentale>.

- process optimization,
- reduced operating costs.

Among the companies that agreed to support one of the case studies was RecordPoint, a leading international records management company actively engaged in AI experimentation, RecordPoint, which accepted to support the case study on NATO Archives declassified and publicly disclosed documents.

The case study on NATO Archives documents provided a significant example of the application of advanced Document AI functions. At the same time, the case study made explicit that these functions delivered the expected benefits only if sufficient data to train the model is available, as stated in European legislation (article 3, paragraph 63). The case study highlighted such a critical issue: even with a very simple (no more than a dozen) single-level category structure, with no internal detailed schema, the margin of error – even after a significant training period – remained significant, and the overall results were quite uncertain, except when contents are easily recognizable.

Another case study analyzed in the study was organized thanks to the involvement of the Regione Emilia-Romagna. The work focused on the unsupervised classification of administrative measures consisting of long, variable texts structured only in the header area using ChatGPT. Also, in this case the classification system used was very simple and limited to the first-level entries of the Regional Council's classification plan, excluding more specific subclasses.

Another interesting case study for the specific objectives of the working group's research has been the one which has concerned CISU (Italian Center for Ufological Studies) study. Even if the case study could not be successfully completed, some useful indications could be drawn from it. The case study was intended to deal with a large series of approximately 100,000 news clippings gathered from publications previously never described in any finding aid: its goal was therefore to explore how AI could help to collect and organize metadata elements – that were formerly not available – on the scanned copies.

In the second phase of the project, specifically in relation to the case study of NATO's declassified and publicly disclosed documents and thanks to the cooperation with other InterPARES studies, another very important aspect emerged, with reference to the significant role that methods and tools for managing documentary heritage play for AI governance: «an opportunity/obligation to document the design and implementation of systems using AI technologies. This perspective is a functional requirement for any significant investment, from project formulation to implementation, including the use of archival document management and digital preservation services»[2].

Such a requirement led us to investigate whether it is possible to confirm, in the course of our experiments, the critical relevance of the role that document management systems are able to play in the planning, implementation, monitoring and validation of AI projects by organizations, in compliance with the documentation and retention requirements identified in the European AI Act[3] .

The goal is to demonstrate that certifying the results and mitigating the risks of any complex project based on AI are achievable and sustainable objectives, if we leverage existing tools that document work processes, projects, and their implementation based on the same rules that have long been the foundation of our document management systems.

As clarified within the working group, we need to move beyond the concept of explainable AI, whose objective is limited to addressing and clarifying the reasons why a given tool produces a certain result

[2] *Ibidem*, p. 133.

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.

starting from a series of elements, and to state instead the principle of maintaining a persistent record of AI project management: this means to create and keep the information and project and process documentation intended to give insight into how a specific technology has been used in a given context and the resulting effects have been achieved. For the above-mentioned reasons, the documentation to be collected and maintained includes, among others, information on the adopted model, the training dataset, and the quality control action, as well as the process data and logs generated by AI tools.

In agreement with the InterPARES study group RP04 responsible for defining a framework for information and descriptive elements necessary to document AI projects, the CU05 study emphasized the need to develop a more standardized structure of information elements which takes into account other standards and recommendations, in particular, the IPELTU[4] (Information Preparation To Enable Long-Term Use) recommendations developed by the CCSDS – Consultative Committee for Space Data Systems (and recently adopted as ISO 23507:2025). IPELTU is meant to define, in compliance with the OAIS model, the category of additional granular information needed to support and document over time complex projects and their crucial phases.

6. Description of the Case Studies

6.1. Preliminary consideration: Paradata Framework for Documenting AI Projects

This section has the aim of describing the methodological framework defined by CU05 for documenting systematically the AI projects. In this context the framework has been applied to the CU05 case study dedicated to the NATO Archives The starting point is the adoption of the principle that building accountable AI requires comprehensive documentation across all phases of a project, covering planning, design, responsibilities, methodologies, processes, decisions, challenges, and monitoring actions. This principle and the related framework (named Paradata Framework) have been developed by RP04 Study “Preserving AI Techniques as Paradata” within the InterPARES ITRUST AI.

The Paradata Framework is aimed to promote transparency and accountability in AI development and deployment and has the capacity to ensure full compliance with European and national legislation on AI use. More specifically, the framework was designed to support practitioners, particularly record managers and archivists, in systematically documenting AI projects, their implementation, and outcomes. This approach reinforces their critical role in preserving the integrity, traceability, and long-term value of AI-related records. Moreover, records managers and archivists are uniquely positioned to develop rigorous methods for managing and preserving records as evidence of responsible and accountable AI use.

To inform the framework for CU05 case studies, the working group conducted a literature review on paradata and additional information elements to determine the types of information required, particularly for the implementation of AI technology when relevant archival functions are in place and must be supported with AI models, such as records classification and metadata enrichment. The literature review led to the establishment of a framework that combines two primary resources: a) the definition of the Machine Learning life cycle (Scott Cameron, 2024), which details the process activities involved in developing and deploying AI models; and b) the definition of the additional information required when complex projects have collected data and have to maintain it over time consistently with the projects life cycle (ISO 23507:2025 - IPELTU). The standard, specifically, outlines the project stages that generate critical information throughout AI implementation activities. For this reason, IPELTU has been considered crucial in the case of AI projects for their documentation.

[4] ISO 23507:2025, Information Preparation to Enable Long-Term Use – IPELTU, freely available as CCSDS Magenta book, <https://ccsds.org/Pubs/653x0m1.pdf>.

The integration of the framework within CU05 study has considered the EU AI Act or Artificial Intelligence Act (2024), with specific reference to the recognition of the role of record-keeping functions in supporting a responsible approach to AI project planning. In fact, the Act highlights the need for:

- strategies: a consistent strategic plan must be in place and preserved;
- responsibilities: roles must be clearly identified across the AI application life cycle, including the level of human supervision;
- inclusion and access: measures must be documented to ensure equitable access and participation;
- transparency: decisions taken and model functionality must be documented;
- data quality and monitoring: accurate, up-to-date data and ongoing monitoring must be maintained according to established policies.

The outcomes of the study, including the methodological framework developed, are available in **Annex 12.1**.

6.2. NATO Archives

NATO Archives was established in 1999 and is based in Bruxelles. Its main scope is the acquisition and preservation of NATO records and the public disclosure of the Alliance records older than 30 years[5]. NATO Archives has made available for case studies a large series of declassified and publicly disclosed Committee records[6] from the Fifties to the Nineties of the Twentieth century. All the documents have been scanned and OCRed and are available in PDF format.

The company providing the technology for the case study is RecordPoint, from Australia. RecordPoint made available its platform called Records365[7]. Records365 is an in-place records management platform. It classifies records to determine their lifetime and will dispose them when disposal is due. Automated classification follows two approaches: 1) expert system (rules-based) classification that uses record metadata, and 2) machine learning classification, based on record text. In Records365 machine learning processes use a supervised learning strategy. Records365 can also use machine learning to categorize records in a way that matches a given classification taxonomy and is able to enrich records by extracting personal identifiers, named entities, and other signals from text and metadata content.

The case study aimed at testing and experimenting with the capabilities of an out-of-the-box application managed by a company specialized in records management products. In particular, the team wanted to test if an AI-enhanced records management tool could perform in an automated way some key records management functions such as records filing and classification, and metadata creation, extraction and enhancement. More precisely, the case-study proposed deliverables were the following:

1. the application can aggregate documents according to clusters, such as archival series/creators/topics-objects/identifier/signatories;
2. the application can capture additional metadata, i.e. signatories, the original security classification of the document, the public disclosure notice, agenda items;
3. the application can flag items that are not NATO documents (i.e., national documents);
4. the application proposes the semantic tagging of the given collection according to controlled vocabulary/ontologies including events/places/people.
5. the application performs text summarization on selected items and/or series of documents.

[5] Nato Archives Online, <https://archives.nato.int>.

[6] NATO Archives publicly disclosed records include the documents of the North Atlantic Council and its sub-committees, the Military Committee and its Working Groups - Committees are created to address specific topics before reporting back to the North Atlantic Council. The fonds and series structure reflects the Committee structure, and its arrangement is based on the reporting chain within the organization and chronologically.

[7] RecordPoint Platform, <https://www.recordpoint.com/recordpoint-platform-tour>.

Automated Classification: Observations and outputs related to Delivery points #1 and #3

With regard to Delivery #1, the research team initially planned to test the application on two distinct document sets: a thematic collection of records characterized by heterogeneous provenance, and a set of Committee document series arranged in simple chronological order. As discussed below, testing could ultimately be conducted only on the latter.

The configuration of the platform for Delivery #1 accounted for the majority of the experimental effort. The implementation followed a standardized workflow for AI-enhanced systems, commonly referred to as the AI application lifecycle. Within this framework, archivists are involved in several key stages, most notably dataset definition, model training, and evaluation.

The process begins with the provision of records ontologies, filing plans, or taxonomies. On this basis, RecordPoint implemented a set of rules within the system reflecting the Archives' conceptual structure. For the purposes of the test, fifteen records series were selected from a significantly larger corpus.

Dataset training represents the second stage of the AI application lifecycle. A minimum of 50 records per category or series is required to train the model effectively. Increasing the volume of training data improves the accuracy of the resulting clustering. Model performance is further refined through user feedback, as corrections introduced during the process are incorporated into subsequent iterations.

Model training is iterative rather than a single, discrete event. Initially, the model is trained on a corpus of documents accompanied by metadata. It is then applied to new sets of documents lacking metadata, which the system automatically assigns to the appropriate series. This is achieved through machine-learning techniques that identify textual references and use them to cluster individual records.

Beyond the initial training set, NATO Archives supplied additional records from the same categories in order to test model performance over time. As the accumulation of records may lead to stagnation in accuracy, periodic re-training is required. Each new training cycle incorporates previously misclassified records, enabling the model to progressively improve through error correction.

Two models were deployed across two testing runs, achieving clustering success rates ranging from 60% to 98% depending on the records series. Variability in performance was influenced primarily by the volume of available training data and by scan quality. The records used in the test date from 1950 to 1990 and are predominantly typewritten. Earlier documents were digitized from microfilm and, in some cases, exhibit reduced legibility, which largely accounts for the observed variation in results.

In parallel, the InterPares research group sought to test RecordPoint's platform on curated thematic collections characterized by non-homogeneous provenance. These collections consist of NATO records originating from multiple series and creators. Testing on such collections proved unfeasible for two main reasons. First, filing plans and taxonomies are either not publicly disclosed or not available, despite their being a prerequisite for platform configuration. Second, the collections did not contain a sufficient number of records belonging to the same archival series to meet the minimum training requirements.

Similar constraints emerged in relation to Delivery #3. NATO Archives holds records with hybrid authorship, such as national records containing NATO-related information, for example, correspondence from ambassadors to the Secretary General. Attempts to test the platform's ability to distinguish between national and NATO records encountered the same limitations identified for curated collections. In particular, the number of available records was insufficient to support model training, and the Archives does not possess filing plans for external records creators, such as national embassies.

Consequently, testing of the platform was confined to the chronological series of Committee documents, for which more than 50 instances per record type could be provided.

Delivery item #1 was therefore completed, albeit within the constraints outlined above. The platform successfully classified records automatically according to a defined section of the filing plan. This constitutes an example of content-based classification, whereby series codes were identified through OCR and relevant classificatory information was extracted. The results were consistent with the hierarchical records structure implemented within the system.

Metadata enhancement: Observations and outputs related to Delivery points #2 and #4

The aim of this delivery item was to assess which additional metadata elements, beyond reference codes, could be automatically captured by the system. In the context of NATO Archives, where security considerations are particularly prominent, one such element is security classification. NATO applies five security classification levels, together with several additional markings. Each NATO document that is publicly disclosed retains its original security classification at the top and bottom of the page and, once released, includes an administrative notation in the left-hand margin referring to the public disclosure notice authorizing its publication.

Testing the automated capture of any of these elements would have required full customization of the relevant metadata fields, followed by model training and re-training, as previously undertaken for archival series classification (see Delivery #1). Instead, the decision was made to test the out-of-the-box metadata enhancement functionality provided by Records365.

The Records365 platform employs an Intelligence Signaling feature to enrich records metadata by detecting signals associated with sensitive information, such as the presence of Personally Identifiable Information (PII). This capability forms part of the platform's broader data intelligence framework, which includes machine-learning-based automated classification. The Intelligence Signaling module supports the identification and protection of sensitive data, including PII and PCI-DSS entities, and is based on Microsoft Presidio technology. It should be noted that NATO Archives records were not used to train the Presidio model and that the records selected for testing do not contain PII.

The PII detection feature was tested on the same set of NATO Archives documents used in Stage #1. The metadata enhancement log reported several false positives. No further investigation was conducted to determine whether calibration of the Presidio model could reduce the incidence of such misclassifications.

Another out-of-the-box capability within the Intelligence Signaling module is the identification of personal names within documents. This feature applies a Boolean metadata tag ('Has person') to indicate whether a personal name is present. While the names themselves are not extracted or stored as metadata, the presence/absence flag proved to be comparatively reliable.

With regard to AI-supported records aggregation, such as entity extraction (e.g. people, places) and semantic tagging based on taxonomies, these functionalities appear potentially applicable to archival contexts. However, at present their adoption remains at a very early stage which requires heavy customization.

Finally, it should be noted that at the time of testing, custom signals within Records365 were available only in beta or early access as part of the Intelligence Signaling module. A maximum of 30 signals per customer was supported, including both out-of-the-box and custom signals. This limitation poses significant constraints when considering the implementation of extensive taxonomies, which are common in archival environments.

Text Summarization: Observations and outputs related to Delivery point #5

The Records365 out-of-the-box feature performs text summarization on selected items and/or document series. The functionality is typically configured for documents written in English; however, as NATO Archives also holds records in French, the workflow incorporates a translation step. Texts shorter than 100 characters are summarized in a single paragraph. Longer texts are processed through ChatGPT, which generates the summary. In the case of French-language records, the text is first translated and subsequently summarized.

The exercise was conducted on only three records due to the costs associated with large-scale implementation. This limitation should be considered by archives considering the use of automated text summarization to support archival description. Constraints are also evident in the quality of the outputs themselves, which are presented in Annex 12.2. Improving the results would require more detailed

prompting, for example by instructing the system to identify elements such as the author, the action, and the object of the document. In this respect, diplomatic analysis could provide a useful methodological framework.

Key conclusions

The way a dataset is defined has a direct impact on the robustness of the results. Consequently, documenting the dataset's biases and limitations — and, by extension, the parameters guiding decision-making — is a critical stage that requires careful attention. Before beginning any project, archivists should select a suitable dataset and evaluate its characteristics thoroughly. Specifically, archivists should consider the potential biases and capabilities of any dataset in relation to:

- *availability*: Is an appropriate dataset available for the intended project?
- *quality and quantity*: Are the records of sufficient digital quality? Are scans clear and OCR-processed? What is the date range and volume of the records? Are there enough records to meet the project's objectives? Are the metadata elements consistently associated with the documents? Insufficient image quality or dataset size will inevitably limit model performance.
- *structure*: Are the records organized according to a clear structure? Are filing plans or ontologies available, and stored in machine-readable or tabular formats? Can this information be shared with external stakeholders? Classification tools must be translated into machine-readable rules that accommodate changes over time. Cross-referencing is also critical: if it is not explicitly defined, the model cannot infer it. Without cross-referencing rules, the AI may struggle to make decisions outside the scope of the predefined rules.
- *language limitations*: RecordPoint currently supports AI/ML classification only in English. However, NATO Archives contains a substantial proportion of records in French, and the system performed reasonably well with these.
- *security concerns*: Can the dataset be shared with external parties? Are there privacy or confidentiality considerations, including the presence of personally identifiable information (PII)?
- *Transportability:* Can the data be efficiently transferred to the AI platform, and by what means? Is cloud storage a viable option? For instance, NATO Archives could share only 14,000 records with RecordPoint out of a potential 300,000 due to limited data exchange capabilities. Tools enabling agile sharing of large datasets are essential prerequisites for projects of this kind.

A further challenge arises when using historical datasets. Archivists are aware that two documents from the same series—one dated 1950, the other 1990—may be simultaneously similar and strikingly different. Such variation can complicate AI training, as models may focus on differences rather than underlying similarities. Ultimately, a model is only as effective as the quality of its input data: the cleaner and more consistent the dataset, the better the model's performance.

Classification

Like many archives, NATO Archives holds a large volume of both digitized and born-digital records. Many born-digital records originate from shared drives with deeply nested structures and limited oversight of their contents. Digitized records come from a variety of sources, and the metadata elements associated with individual documents are often incomplete or inconsistent.

The specificity of the NATO case is, however, generalizable. It reflects situations in organizations where the transition to born-digital records has not been accompanied by consistent file classification, where filing plans have been abandoned or are being phased out, leaving only broad, unspecific categories (*big buckets*). In such contexts, using AI to achieve refined classification objectives—such as records

aggregation or functional classification—is particularly challenging. Simply put, limited structure yields limited results.

For automatic records classification, the precision of the ontology or filing plan provided at the outset directly influences classifier performance. If a highly detailed classification system is in place, very granular training is required; without it, the AI struggles to accurately recognize records, as illustrated by the sub-series example (see Annex 12.2). In some respects, the system performs better with big buckets or less complex classification schemes. Conversely, attempting to use AI for functional classification of elements not already represented in the metadata is unlikely to succeed. AI operates effectively from simple inputs to simple outputs, but not from simple to complex.

A detailed description of the case study is available in **Annex 12.2.**

6.3. Regione Emilia Romagna

The Polo archivistico dell'Emilia-Romagna - ParER (Emilia-Romagna Digital Archival Centre) is a trusted digital repository of the Regione Emilia-Romagna (Italy), responsible for the long-term digital preservation of digital records transferred from all regional public administrations. It functions as a centralized archive to which participating institutions contribute their own archives, benefiting from a high-level professional, archival, technological and organizational service.

It is considered one of the Italian best practices in preservation of digital archives, and its mission is to develop policies and procedures for record-making, record-keeping and record-preservation, to ensure that all data transferred from the public administrations can be safely stored, in order to be accessible in the forthcoming years. ParER preserves different digital record typologies: administrative, educational, cultural, health care, etc. Up to now, it has preserved over 3 billion records, coming from over 100 public administrations.

ParER has its own technological infrastructure, with two physical datacenters (the main one in Bologna and the backup one in Parma). Its services are guaranteed by a totally own preservation software called "SacER", the design, improvement and development of which are under the total control of ParER staff, which amounts to over 20 persons and 15 external professional support.

ParER is also involved in the definition of national models and rules on recordkeeping and digital preservation and participated in various international projects: InterPARES, as partner of the project from 2010, and E-ARK Project as member of archival advisory board.

The study investigated the possibility of automating the assignment of classification classes to administrative documents. The Regione Emilia-Romagna provided approximately 3,000 administrative documents (executive acts and records) for the case study. All documents are original digital files in PDF format. Specifically, it analyzes the experimentation of unsupervised automatic classification at ParER through two distinct research approaches: one conducted in collaboration with the academic field at the University of Bologna, and one carried out internally with support from specialized consultants. The main goal was to evaluate the effectiveness of unsupervised and zero-shot artificial intelligence techniques in assigning classifications to approximately three thousand administrative documents, including executive acts and protocol records, addressing challenges such as the length of the texts, the diversity and inconsistency of content, and the complexity of bureaucratic language.

Given the lack of pre-classified documents required to train supervised systems, the research focused on the feasibility of unsupervised classification techniques. The effectiveness of these methods was subsequently validated by comparing the results with previously correctly classified samples.

The project addressed several complexities intrinsic to the administrative documentation:

- text heterogeneity: long documents lacking standardized structures;
- complexity of the classification plan: a system based on distinct classes of meaning related to different functional areas;
- scarcity of samples: limited availability of homogeneous documents for testing.

The results obtained highlight a clear superiority of Large Language Models over traditional clustering methods or smaller language models. In the academic experimentation, the GPT-4o model stood out for its ability to interpret the semantic nuances of the nineteen classes of a national classification plan model, achieving an accuracy of 65%. At the same time, the internal experimentation conducted on protocol documents showed that the GPT-4.1 model reached a precision of 74% at the first level of classification, while effectiveness dropped dramatically at the lower levels of the regional classification scheme.

The research also demonstrated the value of creating a Silver Dataset generated automatically, which was used to train lighter and more cost-effective proprietary classifiers. These classifiers have, at times, performed better than some open-source models.

In conclusion, although the results of the more advanced models can be considered positive, the research emphasizes that artificial intelligence should be regarded as a critical support for human activity, rather than a replacement, and that archival description tools and essential language simplification are key factors. Further details are provided in **Annex 12.3**.

6.4 Centro Italiano Studi Ufologici (CISU)

The case study involving CISU was not successfully completed, therefore no findings could be derived. However, the report outlines the work carried out in preparation, which may provide useful methodological insights.

CISU (Centro Italiano Studi Ufologici – Italian Center for UFO Studies – headquartered in Turin, Italy) was established in 1985 and possesses the second-largest UFO-related archives in Europe.

CISU accepted in February 2023 to participate in a case study with Team CU05 and make available a series of ca. 100,000 news clippings of publications as a testbed for an AI-based application providing document management services.

Regrettably, the case study was not successful, because it has proved impossible to find a suitable company as a technology partner for the case study. As a matter of fact:

1. a first company selected to support the case study wanted to be paid a substantial amount of money to deliver its services, and its request – according to the project procedures – could not be accepted by the Executive Committee of InterPARES Trust AI;
2. a second company consented to give technological support for free, but after agreeing with Team CU05 and CISU on a set of objectives to be achieved – which included the automatic indexation of contents of news clippings in PDF files processed by OCR, the categorization of the news clippings according to a pre-defined set of types and the identification of duplicates – the company also asked that an appropriate server available for the case-study might be made available for the activities of the case study – or at least a sum of money to rent one – as for various reasons they had none to use for that purpose. After some attempts to find a makeshift solution with the support of the University of Zagreb for providing a server and hosting the research data, the rejection of the request caused the second company to withdraw from the case study, and consequently in August 2024 the case study was definitively shelved without achieving none of the goals that had been set by CISU and Team CU05.

The main takeaway of this unsuccessful case-study is that it is difficult to drive companies with no direct connection with the archives and records management world to work for free or even at a loss for an academic project like InterPARES Trust AI.

It is also important to note that the AI case studies require more technological support and a huge amount of intermediation work than expected in the planning phase.

Further details are provided in **Annex 12.4**.

7. Findings

As a summary of the findings from the overall CU05 study project, this section of the report presents a concise analysis of the functions that AI can support in the field of records and archival management, along with the essential conditions of use that must be observed to maintain quality control over the processes and procedures ensuring the integrity, accuracy, and reliability of documentary and information systems across both public and private sectors.

Table 1 identifies the core functions of records and archives management systems as they have been identified in the first phase of the study, when a survey was carried out in the marketplace[8] : document creation, registration, classification, formation and management of document aggregations, selection and disposal, archiving, and retrieval. For each of these functions, specific activities in which AI supports users are presented, as well as requirements to be met to ensure the quality of the results.

This analysis highlights that AI can effectively support several key functions in records and archives management, including records and content retrieval, data extraction, text and summary generation, description enrichment (also in multilingual environments), personal data protection, automatic or semi-automatic classification, and indexing, but only if and when specific conditions are considered and implemented.

AI-driven data and description enrichment produces reliable results when applied to well-defined object types, coded and standardized markers, and a sufficient number of instances. Personal data protection proves effective only when the legal context and the categories of personal data are clearly identified, with risk assessment serving as an essential safeguard. Automatic classification can produce useful results when classification/file plans, structured guidelines, specific workflows, and clearly characterized content or communication channels are in place. However, substantial human involvement in training and evaluation remains necessary, highlighting that AI applications in the archival field are still at an early stage and require significant investments in both costs and resources.

Table 1. AI core functions of records and archives management systems

	Supported activities	Requirements and Checks

AI for Records Creation	– Drafting documents
	– Automatic text completion
	– Generating abstracts and summaries
	– Writing assistance
	– Personalization
	– Automatic translation
		– Careful legal risk assessment and consideration of:
		  – ethical content use
		  – compliance with intellectual property rights
		  – compliance with privacy regulations
		– Rigorous training and verification to ensure content quality and consistency, preventing AI tendencies toward generalization and oversimplification, particularly in scientific or technical texts[9]

[8] Allegrezza S., Guercio M., Mata Caravaca M., Grandi M., La Sorda B., "CU05 The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas – Vendor Survey. Final Report", InterPARES Trust AI Research Document, 1 November 2023, page 25, <https://interparestrustai.org/trust/research_dissemination>.

[9] U. Peters, B. Chin-Yee (2025). Generalization bias in large language model summarization of scientific research, Royal Society Open Science 12 (4), <https://doi.org/10.1098/rsos.241776>.

	Supported activities	Requirements and Checks

AI for Records Registration	– Automatic data recognition and extraction
	– Data validation
	– Automatic registration
	– Provenance recognition
	– Automatic document assignment (e.g., emails)
		– Careful analysis and rigorous training on critical information:
		  – document subject (accuracy, privacy concerns)
		  – correspondents: senders and recipients (correctness and completeness)
		– Human verification and validation at all stages of planning and implementation

AI for Automated Classification	– Content-based classification (managed with OCR) – Extraction of information useful for classification – Creation or compliance with hierarchical structures	– Rigorous selection and verification of supervised learning methods, especially for complex and articulated classification systems – Requirement for well-defined structured information, ideally through classification/file plans and supporting documents – High-level human oversight to ensure consistency in classification operations and hierarchy creation
AI for Records Aggregation and Management	– Entity extraction (names of people, places, institutions; event types) to facilitate functional aggregations – Analysis of complex functional contexts related to classification/file plans and organizational charts – Automatic document grouping guided by algorithms based on formal similarity criteria	– High-level human monitoring of the functional aggregation criteria and their consistency with the classification/file plan, taking into account the legal significance of the archival bond that governs the creation of records aggregations, particularly in the case of procedural, activity, and business files
AI for Selection and Disposal	– Planning selection and disposal – Automation of disposal processes and related documentation through process workflows and reporting templates	– Availability of an accurate and complete retention plan and classification/file plan for document aggregations, from which necessary information can be derived – Definition of forms and reports consistent with legal obligations and internal and external decision-making processes

	Supported activities	Requirements and Checks
AI for Archiving and Retrieval	– Generation of relevant metadata for preservation, description, and retrieval across the records lifecycle, including records centers and historical archives – Creation of ontologies and formal descriptive models to enable semantic search – Full-text search with complete indexing – Advanced search using visual and semantic comparison tools, suggesting related searches – Exploration of the archive using information on its hierarchical structure – Support for preservation planning in relation to regulatory and fiscal obligations, as indicated in the retention plan – Monitoring format and media obsolescence	– AI can handle complex meanings and perform relevant searches provided qualified metadata and/or domain ontologies exist to limit inevitable ambiguities – AI agents can interact with users during query and search phases if training has accounted for contextual information (classification/file plans, organizational charts, directories) – An accurate and complete retention plan is required for preservation planning

The study also emphasizes that these functions are subject to several critical constraints that must be managed by qualified professionals. These include the inherent complexity of documentary heritage, the indispensable role of human mediation and labor-intensive verification processes, the significant and ongoing investment required for model training, the rapid evolution of AI research and application environments, and the currently limited AI-related expertise among records and archives management professionals.

8. Conclusions

8.1. Remarks on classification, aggregation and indexation of records

The exam of the various features of AI and the ways it has been used to support archives and records management has made it possible to identify the prospects and issues currently existing as far as the use of AI for archives and records management is concerned.

Regarding the classification, aggregation and indexation of records, AI is potentially capable of assisting professionals in dealing with various labor-intensive tasks which often cannot properly be carried out because of lack of sufficient human and technological resources.

As far as classification is concerned, if AI can rely on appropriate workflows, sufficient metadata elements and recognizable features, the outcomes may be positive and able to help to sort automatically records according to the categories of a file plan. As the enormous number of records that are daily produced often prevents proper classification and puts a huge strain on the human resources available for this activity, it is understandable that the coming of AI has raised great hopes in this respect and could play a significant role in the future.

Similar expectations exist in relation to the description and indexation of records, with specific reference to the capacity of AI to enrich archival descriptions and autonomously find metadata elements to be associated with records and aggregations of records, on condition that AI is aptly trained and may be able to unequivocally identify the objects that are to be described, apply relevant standards and use codified metadata elements.

Likewise, AI might also help in the task of aggregating records in casefiles, series, volumes and any other kind of grouping to organize records based on common and meaningful features and the respect of the same conditions implied in the classification activities. As a matter of fact, the activity of describing records and aggregations easily requires vast amounts of resources, and AI might meet the needs of the archival community in relation to this task. Moreover, AI may perform this job by using a wide array of languages, so as to quickly produce multilingual indexes and finding aids.

However, it is to be considered that the specific nature of records and of the information they contain and convey is even more complex than those of other information objects, as usually you can gain intelligence over a record or an archival aggregation only if you are able to understand the relations existing between different records and between single records and the aggregations they belong to: this presents challenges to humans and to AI-based applications as well.

8.2. Market survey and case studies: main achievements

The findings from the survey conducted in 2022-2023 on the companies producing AI-based applications of interest for archives and records managers have shown that at this stage it is still difficult for present AI technologies address several aspects for the records and archives management: although, seemingly, almost all of the companies participating in the survey offered automatic classification, when we had asked whether they were also able to provide other services specific to archival discipline - such as aggregation of records to casefiles or identification of the provenance - the number of the companies which could deliver the services has significantly decreased; moreover, even those that had replied in the affirmative have often detailed some conditions for supplying these kinds of services: e.g. “If there is metadata to represent those processes (e.g. a case file number)”[10] or “If the involved entities are stated in the content of the document and the type of relation linking them is linguistically expressed”[11] .

Then again, also when a company had answered they could perform a particular kind of service, it has always been essential – especially in case of companies not possessing expertise in archival sciences - to verify whether what was on offer actually matched the concept as defined from the archival perspective: this is particularly true for classification, a sort of catch-all term sometimes pointing to actions having no connection with archival discipline.

During the second stage of its activities, Team CU05 has been able to focus more closely on how the specific needs of archives and records management could be met, through case studies where researchers have interacted with the developers and project managers who have built and operated particular AI applications.

Overall, the findings from the case studies point out that archivists and records managers might derive significant benefits from the application of AI-based technologies, but that also a huge amount of work is required to reap them.

[10] Allegrezza S., Guercio M., Mata Caravaca M., Grandi M., La Sorda B., “CU05 The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas – Vendor Survey. Final Report”, InterPARES Trust AI Research Document, 1 November 2023, page 25, <https://interparestrustai.org/trust/research_dissemination.>

[11] *Ibidem*.

The case studies have clearly revealed the great deal of effort to be put in any project involving the use of AI for records management: this is true in the preliminary stages as well as in the execution of the project and in all the auditing activities.

With reference to the case-study which has involved NATO (where it is be noted that the IT company that has provided the AI-based technology has a solid archival background), it has sometimes been difficult to supply the developers with sufficiently large test-beds for the training of the AI application (a minimum of 50 documents per each category of the classification scheme was required – a condition sometimes hard to be met), while the presence of multiple and seemingly self-conflicting (for the AI application) metadata elements have made it necessary to refine several times the model underpinning the AI application.

Also the case-study organized by Regione Emilia-Romagna in Italy has highlighted how much preparatory work has been required firstly just to understand the better technique to build an AI-based tool to categorize current records according to the classification scheme used by Regione Emilia-Romagna; secondly – after the decision to adopt an unsupervised zero-shot classification approach, where the AI model is given only a natural language description for each class – to create a dataset of records to train the AI model, by being careful to select a sample representing all the branches of the classification scheme; and finally to refine and audit the AI model that had been built.

The sheer number of resources and effort to be put into action and the steadfast work necessary to monitor and refine the AI solution that has been chosen may prove to be disappointing to everyone who had hoped to find a shortcut to quickly reduce the workload entailed by records classification, aggregation and indexation. At the moment AI-based technology does not let us achieve quick wins in reference to such tasks: specific organizational and technical conditions are required, and remarkable levels of commitment and human intermediation are essential to attain good – or at least meaningful – results.

No AI powered tool can yield good outcomes if the context where it is used is flawed or poorly organized. As far as archives and current records are concerned, if an AI application cannot rely – e.g. – on proper metadata elements, sufficient contextual information, adequate intellectual tools and finding aids, it is unavoidable that what can be carried out by using the AI application will possess a shoddy quality level even in the best case scenario (while in the worst case scenario we might be confronted even with gross mistakes and serious non-compliance). AI technology may be of use only if humans enable it to work properly. Therefore, it is imperative for information professionals to remind openly everyone that any deployment of AI-technologies cannot replace good archives and records management practices, which AI is supposed to support, but definitely not to displace them (bearing in mind, moreover, that AI-based applications are trained by documents and materials kept in the archival and document holdings they are expected to help to manage – we can easily guess the consequences of a training performed on misleading testing grounds…).

The outcomes of the case studies concerning NATO and Regione Emilia-Romagna – with reference to classification tasks - have evidenced that at the moment AI-based applications are still a far cry from being tools that can be used without strict human supervision:

1) Regione Emilia-Romagna, after testing various approaches and fine-tuning its AI-based tool, during a trial session manage to achieve – by using a zero-shot classification strategy and GPT as a Large Language Model – a 65 % accuracy level in the classification of administrative acts, while in another test involving this time all the kinds of documents produced or received by Regione and carried out again through GPT the exact match rate with the right classification was hit in the 74% of the cases (but 74% only for correctly matching the 1st tier subdivisions of the file plan; for the lower tiers of the file plan the exact match rates have been significantly lower)[12];
2) In the NATO case-study it was not possible to attain the threshold of 80% of accuracy for all the classes of records even during the development and training of the AI model, while the testing of the

[12] See the outcomes achieved as shown at Annex 12.3, paragraph 5 Conclusions, table 23.

model over a body of records to be classified has shown that the prediction probability rate calculated by the AI model for each proposed classification (i.e. the “confidence” of the AI application on the classification it itself had proposed) was often rather limited[13].

These upshots imply that at the present time very strict control by humans over the results given by these AI tools is essential, as the outcomes of the case-studies show that it is extremely difficult for an AI-based application to match the success rate of a knowledgeable human operator. It is quite possible that with optimal conditions and a longer and more specific training an appropriately developed AI application may match or even exceed the success rate of an experienced human in records classification – if not now at least in the future; but for the time being that appears to be a hard target to hit.

Finally, it is crucial to be able to document any stage of the process through which AI tools have been planned, developed, deployed and implemented, for both transparency and efficiency reasons:

- for transparency, as the AI-based applications used for archives and records management will be used to handle work processes heavily impacting on the lives and entitlements of human persons and on public institutions; it is therefore imperative to make available pieces of evidence which may enable us to reconstruct and understand as much as possible the whole process through which an AI-based application has performed its tasks, and the more the work of the AI-based application impacts on fundamental rights, the more compelling this requirement is.
- for efficiency, since situations where unexpected outcomes and kinds of behaviour of an AI-based application create problems are realistic and sometimes have already occurred: to have at hand a body of knowledge through which we can explain both the reasons for given results and the inner workings of an AI-based technology is therefore also an important business need.

8.3. The information framework for AI project and the role of recordkeeping

With reference to the transparency and the accountability, Team CU05 researchers have tried to build a first framework to identify the specific pieces of information that may be relevant to detail all the stages of an AI project, and have also tried to use the elements of the framework in the NATO case-study: although the first application of the framework has been difficult for various reasons – not least because the IT company taking part in the case-study had never been asked to provide these types of information –, the set of information that has been gathered represents an initial step to establish a body of knowledge crucial to achieve the above-mentioned objectives of transparency and efficiency.

Of course, we have to be aware that in some cases particular kinds of information important or even essential to document an AI-based project might not be available, because e.g. the provider is unable to make it available or the disclosure of some information might be a breach of trade secrets, but it is paramount to know which information are relevant for to give evidence of how an AI-based application has been developed and implemented: in some scenarios, the fact that we cannot access and obtain specific information might be an important aspect to be considered in a risk assessment and might also lead to the dismissal of the application itself, in case the lack of data should make it impossible to account for how an AI-based product with potentially serious impacts on the lives of people and fundamental rights has been built and used.

Then again, various legislative environments are starting to enact laws and regulations concerning AI (e.g. EU AI Act, Colorado AI Act), and the first relevant standards have been published (e.g. ISO/IEC 42001). It is evident that adequate proof of compliance with what each of these sets of rules dictates is required, and that in turn makes it necessary to prepare the information needed to substantiate such proof.

[13] See what has been stated at the Annex 12.2, paragraph 4 Evaluate AI Model - Planning Phase and the outcomes achieved as shown at the figures 5, 6 and 7 of Training AI Model - Closing Phase

8.4. Final recommendations

We suggest, to conclude, a series of main takeaways that might also help to assess whether and how to use or not a specific AI-based application in a particular project or circumstance:

- Human intermediation is key to any stage of the process, from planning to training, from deployment to execution, from monitoring to audit. This is not a stand of principle but is a fact that Team CU05 researchers have been able to verify during the case-studies of the project. Not only is human intermediation indispensable for the success of any project or program involving AI, but it is almost always necessary to allocate sizable amounts of resources (including, of course, the time and effort of human workers) to carry it out successfully.
- The classification, aggregation and indexation of records – in the way as they are understood by archivist and records managers – are activities still rather complex for AI-based technologies, and this is particularly true for the aggregation of records in case-files and some particular tasks such as appraisal or the identification of provenance: it is therefore extremely important to set carefully the objectives that are to be achieved, since – if it may be impossible to foresee exactly all the consequences of the use of an AI-based application – we can at least establish goals, benchmarks and quality levels that will help us to take stock of the situation at different moments during the execution of a project or a program; it is likewise essential that the expectations of all the stakeholders may be adequately managed, so as to avoid that the pervasive hype surrounding any kind of AI-based technology – and often fostering both baseless fears and unrealistic hopes – may have a detrimental impact on the advancement and outcomes of the project or program.
- In a strict connection with the former points, it is necessary to assess attentively the conditions of the materials or systems which are to be processed by AI-based technology. If – e.g. – the training of an AI tool has taken place by using sets of data that are inconsistent, of poor quality, full of wrong information or even maliciously tampered – and that because the shoddy state of the document holdings does not let you provide the AI tool with a better kind of data – it goes without saying that the outcomes of the actions performed by the AI tool will be poor or in some cases even harmful. If – e.g. – a classification scheme is too complex and has categories which are not properly distinct, AI-based applications in any case will have trouble assigning the correct classifications to documents (in the same way as humans will).
- As always – but even more so in the case of AI technology –, to avoid misunderstanding and disappointment, archivists and records managers must fine-tune their professional terminology with that of other professionals or business roles that may use the same word to express a concept completely different from that acknowledged in the archival domain: just to give an example, Team CU05 researchers have appreciated that the expression “Intelligent Document Processing” (IDP) has often little to do with the “document processing” as understood by archivists and records managers.
- It is essential to document every step of the project or program where the AI-based application is meant to be used. The kinds of information that are listed in the paradata information framework integrated by Team CU05 might be a good starting point.

AI might surely be a great opportunity to improve the classification, aggregation and indexation of records, but only provided that we shall be able to ensure and create the conditions to enable it to work for the benefit of archives and records management, and of the public at large.

9. Dissemination

9.1. Publications

Allegrezza S., Mata Caravaca M., Grandi M., Guercio M., La Sorda B. (2024). *The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas*. L. Duranti and C. Rogers eds. SCEaR Newsletter. Special issue. Artificial Intelligence and Documentary Heritage, , pp. 43-48, https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf

Guercio M., Mazzeo A. (2025). Intelligenza artificiale nel trattamento dei documenti e degli archivi. *Tempo presente*, 2025, 1, pp. 94-107 https://tempopresenterivista.altervista.org/e-uscito-il-numero-529-531-di-tempo-presente-dedicato-monograficamente-al-tema-focus-intelligenza-artificiale-lenigma-della-singolarita-e-le-leve-umane-per-la-conoscenza/

9.2. List of deliverables and conferences

Magnoni F. et al. (2025), *The role of Artificial Intelligence in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas. A Practical Case Study and a Project Reporting Framework*", ICA Congress, Barcelona, 27-29 October 2025 https://icabarcelona2025.cat/programme/

Allegrezza S., Caravaca M. M., and Magnoni F. (2025), *AI and Classification: Case-study and paradata framework*, "AI for Good - Public Records and Artificial Intelligence" Conference, Bologna (Italy), June, 20, 2025.

Allegrezza S. (2025), *Classificare e fascicolare i documenti con l'AI: l'esperienza del progetto InterPares Trust AI*. Webinar Formez "L'intelligenza artificiale negli archivi: soluzioni in campo e casi concreti", June, 10 2025.

Allegrezza S. (2025), *Classificare e fascicolare i documenti con l'intelligenza artificiale: luci ed ombre", Conference "Archivi e intelligenza artificiale responsabile. Esperienze, opportunità, cautele*, Bologna. May, 16 2025

Allegrezza S. (2024), *L'impiego dell'intelligenza artificiale per la ricostituzione delle aggregazioni archivistiche e l'arricchimento dei metadati negli archivi digitali*, XII convegno nazionale AIUCD 2024, Catania (Italy), 28-30 May 2024.

Allegrezza S. (2024), «Il ruolo dell'intelligenza artificiale nella creazione o ricreazione di aggregazioni archivistiche di documenti digitali e nell'identificazione di schemi di metadati», ForumPA 2024. Rome, May, 23 2024

Allegrezza S. (2024), The role of artificial intelligence in creating or recreating archival aggregations of electronic documents and identifying metadata schemas, XXX International Scientific and Practical Conference "Documentation in Information Society"», Moscow, 18-19 April 2024.

Esteval Casellas L. (2024). Retos y oportunidades de la Inteligencia Artificial para los archivos, Conferencias Archísticas Pontifica, Universidad Católica del Perú (PUCP), April 16, 2024, https://www.facebook.com/ArchivoPUCP/videos/856371329585841

Esteval Casellas L. (2024). *Intelligentia ex machina. Reotes I oportunitats de la Intelligencia Artificial als Arxius*, VII Jornada Valenciana de Documentació. RESET. Intelligència artificial: nous paradigmes en gestió de la informació, March 2024, 1, https://www.youtube.com/watch?v=Jq9-bb9CGTA

Magnoni F. (2024). *The role of Artificial Intelligence in identifying or reconstituting archival aggregations of digital records and enriching metadata schema*, 49th ICA-SIO Annual Meeting & Workshop 21-23 May 2024, The Hague-Netherlandshttps://www.ica.org/event/49th-sio-annual-meeting-workshop/

Magnoni F. (2024), *The InterParesTrustAI Project and the role of Artificial Intelligence in identifying or reconstituting archival aggregations of digital records and enriching metadata schema*, NATO Archives Committee, 18-19 June 2024, Bruxelles-Belgium

Allegrezza S. (2023), *Una prima analisi delle piattaforme di Artificial Intelligence per la salvaguardia del vincolo archivistico: lo studio CU05*, Workshop «Il progetto I TRUST-AI e il contributo della comunità archivistica italiana» Rome, November, 17 2023.

Esteval Casellas L. (2023). Un modelo de anàlisis funcional, September 2023, Presentation at XXVIII Jornadas de Archivos Universitarios (CAU-CRUE), https://youtu.be/WB-3Z9VLsRk?t=1076

Musa M. (2023). *I Trust AI CU05. The use of AI in identifying or recreating archival aggregations: the case of NATO archives and use of AI technologies developed by RecordPoint*, Workshop on AI in the Archives, Luxembourgh, European Parliament (Directorate for Innovation and Central Services Attached to the Secretary-General Archives Unit), 16 November 2023, https://docs.google.com/presentation/d/1f4htpWQhTOI1u7I6ty4EAUU2sRfE-W79/edit#slide=id.p1

10. Further research

Some topics here analyzed are very crucial at European level specifically regarding the application of the Regulation (EU) 2024/1689 (Artificial Intelligence Act) which implies the obligation of documenting AI projects and keeping/archiving the related information. For this reason an initiative with the aim of exploring a standardized use of information (based on paradata framework and Ipeltu standard) for documenting AI projects and preserve them for future analysis has been approved in February 2026 by the Italian standardization body (Ente italiano di normazione – UNI) and will be proposed to the European Committee for Standardization – CEN, in particular to the Technical Committee 468 Preservation of Digital Information.

11. References

11.1. Standards

DoD 5015.02:2015, Incorporating change 2017, Design Criteria Standard for Electronic Records Management Applications
https://www.dmi-ida.org/knowledge-base-detail/dod-instruction-5015-02

ISO 15489:2016 Information and documentation - Records management

ISO 23081-1:2017 ISO 16175-1:2020 Information and documentation — Processes and functional requirements for software for managing records – 1. Principles

ISO 30301:2019 Information and documentation - Management systems for records – Requirements

ISO 16175:2020 Information and documentation - Processes and functional requirements for software for managing records

ISO 23507:2025 Space data and information transfer systems — Information preparation to enable long term use (IPELTU). CCSDS Magenta Book
https://ccsds.org/Pubs/653x0m1.pdf

ISO/IEC 42001:2023 Artificial Intelligence management systems

Moreq - MOdel REQuirements for the management of electronic records, 2010.
https://moreq.info/

11.2. Legislation

Colorado Artificial Intelligence Act - CAIA, Senate Vill 24-2025

Regulation (EU) 2016/679 - Protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

Regulation (EU) 2024/1689 – Artificial Intelligence Act,
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

11.3. Literature

Cameron, S., Hamidzadeh, B. (2024). Preserving paradata for accountability of semi-autonomous AI agents in dynamic environments: An archival perspective https://ssrn.com/abstract=4681230 or http://dx.doi.org/10.2139/ssrn.4681230).

Franks, P., Davet, J., Hamidzadeh, B. (2023). Archivist in the machine: paradata for AI-based automation in the archives. *Archival Science*, 2023, 23:275–295, https://doi.org/10.1007/s10502-023-09408-8)

Giovannini M.P. (editor) (2025). *Intelligenza artificiale e archivi digitali. Interazioni e governance.* Padova: Collana di Minigrafie, Associazione Siav Academy.
https://www.associazionesiavacademy.it/it/intelligenza-artificiale-gestione-documentale/.

Peters, U., Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Open Society open science*
https://royalsocietypublishing.org/doi/epdf/10.1098/rsos.241776.

12. Annexes

12.1 Paradata Framework for Documenting AI Projects

The study initiated conducting literature review on paradata and additional information elements to determine the information required for documenting the CU05 case studies. These case studies focus on implementing AI technology in archival functions, especially in records classification and metadata enrichment. These paradata or additional information elements refer to data that provide insights into the process and methodology used in AI implementation, as well as the decisions and challenges faced throughout the case studies. This documentation is essential for assessing the effectiveness and impact of AI integration into archival practices, as it not only promotes transparency and accountability but also provides practitioners with a structured framework to preserve evidence of responsible AI use in archival contexts.

The literature review led to the establishment of a framework that combines two primary resources:

a) the Machine Learning life cycle (Scott Cameron, 2024), which encompasses the process activities involved in developing and deploying AI models, such as:
    1. Identification and preparation of datasets;
    2. Producing AI model;
    3. Training AI model with datasets prepared;
    4. Evaluating AI model performance;
    5. Implementing AI model;
    6. Improving AI model with new data;
    7. Monitoring operations;
b) the phases of data collection (ISO 23507:2025 - IPELTU), which outlines the project stages that generate critical information throughout the activities involved in implementing AI. These phases are:
    1. Activity Initiation and Planning: Establishing the project's objectives, timelines, and requirements;
    2. Activity Executing: Carrying out the activities involved in developing and deploying the AI model;
    3. Activity Closing: Concluding the project once all requirements, phases, or contractual obligations have been completed.

Following the initial definition of the Paradata Framework (Table 2), the study elaborated on the specific types of information elements to be incorporated within each framework cell.

The framework is organized detailing the lifecycle or stages of a document management–related AI project listed in the first column, and the three phases associated with each stage – Planning, Executing, and Closing – presented in the first row. The cells are populated with categories of information elements that capture the environment, development, features, and organizational and technological processes of AI systems handling documents (Table 2). For clarity, we refer to this set as "processing information elements," a term inspired by the <processinfo> element of the EAD ("Encoded Archival Description") standard, though in this context the scope of processing activities is considerably broader.

For the specification of the kinds of suggested categories of such information elements, we have considered the following resources:

1. Specific Standards or Recommendations advising the collection of information elements whose characteristics make them eligible to be viewed as example of processing information elements, such as IPELTU (Information Preparation To Enable Long-Term Use) Recommended Practice

CCSDS 653.0-M-1, issued in 2024 by The Consultative Committee for Space Data Systems, especially chapters 4.2.2 and 4.2.3 [currently, ISO 23507:2025]; and Moreq ("MOdel REQuirements for the management of electronic records") 2010, as far as the categories of logs reported in the table are concerned.

2. The indications found in the relevant literature concerning paradata with specific attention to the articles written by Scott Cameron, Babak Hamidzadeh, Patricia Franks, Jeremy Davet and Jenny Bunn from 2022 to 2024, in particular "Archivist in the machine: paradata for AI-based automation in the archives" (P. Franks, J. Davet, B. Hamidzadeh "Archival Science", 2023) and "Preserving paradata for accountability of semi-autonomous AI agents in dynamic environments: An archival perspective" (S. Cameron, B. Hamidzadeh, 2024).
3. Professional knowledge derived from the expertise of the members of Group CU05 members who have managed projects and programs involving digital archives and records management. For instance, this is why risk assessments were included as a possible category of processing information elements, and why the category "Statement about the quality levels expected by the AI model" (point 4.1.1) was suggested. Both risk assessments and statements regarding expected quality levels are commonly used in corporate projects.

Given the preliminary nature of this initial paradata framework, which requires further testing with case studies and discussion within the InterPARES community, we have deliberately limited our proposal to the main categories of processing information elements. A more detailed and granular list, including homogeneous subsets or individual elements, would need to be developed following further verification processes, ensuring that the framework accurately reflects practical applications and documentation needs.

Table 2 - Structure of the Paradata Framework

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
1. Identification and preparation of datasets			
2. Produce AI model			
3. Train AI model			
4. Evaluate AI model			
5. Implement AI model			
6. Improve AI model			
7. Monitor operations			

Table 3 - Structure of the Paradata Framework with the main categories of information elements

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; Adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
1. Identification and preparation of datasets	1) Information about the purposes and goals to be achieved by getting the datasets processed by AI model 2) Information about the selection and preparation of the datasets 3) Information about the agents (e.g. organizations, other IT systems, people, public authority) that have made available the datasets 4) Information on the constraints (deriving from copyright, privacy, contractual, other legal, technology issues, etc.) bearing on the datasets	1) Information about the datasets to be processed by the AI model (structure; type of data of the datasets; history of the datasets) 2) Information about the control, cleaning and adjustment of the datasets 3) Logs of any operation performed on datasets	1) Outcome of the quality checks performed on the datasets to be processed by the AI model 2) Compliance statements (e.g. statements where the fact that the datasets are deemed to be fit for the purpose is attested to; statements that the datasets to be processed comply with the relevant legislation) 3) Information on the roles and offices involved in the identification and preparation of the datasets

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; Adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
2. Produce AI model	1) General design and planning documentation to produce the AI model 2) Information (Designs, plans, functional analysis maps) about the design of the algorithms underpinning the AI model 3) Information on the constraints (deriving from copyright, privacy, contractual, other legal, technology issues) bearing on the production of the AI model 4) Methodologies followed to design the AI Model 5) Information about the variables and constants expected to bear on the way the AI model works 6) Information about the roles and offices involved in the production of the AI model 7) Risks assessments concerning possible problems intrinsically relating to the design of the AI model	1) Information about the software actually used to produce the AI Model (including - if possible and applicable - the source code of the software) 2) Values of variables and constants actually used to produce the AI model 3) Progress reports describing the development of the AI Model 4) Information about contracts drawn up with and activities of consultants and suppliers involved in the production of the AI model 5) Logs of the operations conducted to develop the AI model	1) Versioning general information (i.e. versioning specification, detailing the sequence number of the version, who is or are the owner(s) etc.) about the first version of the AI model 2) Outcome of the final review and quality checks performed on the intermediate and final versions of the AI model 3) Compliance statements (e.g. statements where the fact that the AI model is deemed to be fit for the purpose is attested to; statements that the AI model complies with the relevant legislation)

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; Adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
3. Train AI model	1) Design of the activities carried out to train the AI Model 2) Methodologies followed to train the AI Model 3) Methodologies followed to find and build the training datasets 4) Information on the roles and offices involved in the activities carried out to train the AI Model 5) Information about the agents (e.g. organizations, other IT systems, people, public authority) that have made available the training datasets 6) Risks assessments concerning possible problems intrinsically relating to the organization of the training of the AI model	1) Information about the training datasets actually used to train the AI Model 2) Information about the control, cleaning and adjustment of the training datasets 3) Logs of the training activities 4) Progress reports describing the activities performed to train the AI model and the outcomes of the training	1) Final report about the outcome of the training (or of a stage of the training) of the AI model 2) Outcomes of the quality checks performed on the training datasets 3) Compliance statements (e.g. statements to assess whether the processes deployed to training the AI model have been adequate or not)

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; Adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
4. Evaluate AI model	1) Statement about the quality levels expected by the AI model 2) Information about the methodologies and metrics used to evaluate the AI model 3) Information about the design of the tools used to evaluate the AI model 4) Information on the roles and offices involved in the evaluation processes 5) Risks assessments concerning possible problems intrinsically relating to the the way the AI model has been created	1) Logs of the measurements and controls carried out to evaluate the AI model 2) Reports about any evaluation derived from any event (e.g., report about an egregious bias manifested by the implementation of the AI model) connected the implementation of the AI model 3) Logs of the tools used to carry out the measurements and controls	1) Reports on the outcomes of the quality checks performed on the capabilities shown and the work done by the AI model 2) Reports on the outcomes of the quality checks performed on the evaluation tools and processes used to evaluate the AI model 3) Compliance statements (e.g. statements to assess whether the processes deployed to evaluate the AI model are adequate or not) 4) Periodical reports evaluating the performance of the AI model

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; Adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
5. Implement AI model	1) Information about the expected behaviour and expected outcome of the AI model 2) Risks assessments concerning possible problems that could emerge during the implementation of the AI model 3) Information about emergency procedures to be implemented to solve possible problems relating to the implementation of the AI model (e.g. sudden interruption of services; damages to rights of people affected by the behaviour the AI model, etc.) 4) Information on the roles and offices involved in the implementation processes	1) Logs of the actions and decisions made by the AI model 2) Logs of any event which has or may have impacted on the AI model (e.g., actual or possible disruptions of the business continuity; scheduled updates; actual values of the variables and constants not produced directky by the AI model but assumed by the AI Model duriung its operations; etc.) 3) Logs of the interactions between the AI model and other IT systems (irrespective whether they are AI-based or not)	1) Periodical reports about the outcomes of the implementation of the AI model 2) Reports about any issue or meaningful event which has taken place during the implementation of the AI model 3) Periodical reports about the behaviour of any IT system interacting with the AI model

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; Adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
6. Improve AI model	1) Information about the rationale to carry out modifications to the AI model 2) Information about the design of the changes to be developed to modify the AI model 3) Information about the expected outcomes, benefits and risks deriving from the modifications carried out on the AI model 4) Information on the roles and offices involved in the modifications to the AI model 5) Risks assessments concerning possible problems relating to the proposed improvement of the AI model	1) Information about the development activities carried out to modify the AI model (including - if possible and applicable - the source code of the software) 2) Logs of the modification activities 3) Description of any modification actually carried out to the AI model	1) Versioning general information (i.e. versioning specification, detailing the sequence number of the version, who is or are the owner(s) etc.) about the new version of the AI model 2) Outcome of the final review and quality checks performed on the modified version of the AI model 3) Compliance statements (e.g. statements where the fact that the modifications to the AI model are deemed to be fit for the purpose is attested to; statements that the modifications to the AI model comply with the relevant legislation) 4) Follow-up reports describing the results of the new version of the AI model

INFORMATION FRAMEWORK FOR DOCUMENTING AI PROJECTS			
(Based on the AI application life cycle and project phases; Adaptation by InterPARES Trust AI, CU05 Study, 2025)			
Lifecycle of AI Application (From: InterPARES Study Trust AI, RP04 Study 2024)	**Phases of Data Collection** (From: IPELTU 2025)		
	1. Planning	**2. Executing**	**3. Closing**
7. Monitor operations	1) Rules, procedures and tools designed to monitor the AI model 2) Information about the methodologies and metrics used to evaluate the AI model 3) Information on the roles and offices involved in the monitoring processes	1) Information about any deviation from the way the AI model was expected to work (i.e., from the initial planning of the AI model) 2) Logs of the monitoring operations 3) Logs of the tools used to carry out the monitoring operations 4) Claims and protests raised by anyone affected by the implementation of the AI model 5) Logs of other systems (irrespective of whether they are AI-based or not) which work in dependency on the AI model or are in any way affected by it 6) Reports about any maintenance action carried out to support the implementation of the AI model	1) Periodical reports produced to describe the monitoring operations and their outcomes 2) Compliance statements (e.g. statements where the fact that the monitoring operations of the AI model are deemed to be fit for the purpose is attested to) 3) Information about activities carried out to ensure the long-term preservation and availability of the monitoring operations

12.2. Information Framework for Documenting AI Projects: NATO Archives / Record Point Case-Study

The following pages present a test application of the Paradata framework in the context of the NATO Archives–RecordPoint case study. The exercise contributed to the identification of additional issues related to roles and responsibilities in the documentation phase, particularly concerning the respective tasks of archivists and software providers. The purpose of this section is to examine how an AI project can be documented from a practitioner-oriented perspective and to assess the applicability of the paradata concept within a real-case scenario.

From this perspective, the analysis is conducted from the standpoint of an archivist seeking to document the processes involved in the implementation of an AI-enhanced recordkeeping solution. More specifically, the aim is to map the information collected through the case study against the identified phases and stages of the AI project. In addition, the various stages of the project are described on the basis of the available information and related processes.

It should be noted that, as this analysis was conducted ex post, some of the “additional information” identified by the framework was not available in relation to the case study under consideration.

For the purposes of project description and documentation, the most up-to-date materials available from the case study were identified and collected. These include, among others, the initial project proposal, a log of the records used in the study, the logs generated by the machine learning model exercises (where available), and the machine learning report provided by RecordPoint.

The case study aimed to assess the capabilities and requirements involved in implementing an AI-enhanced recordkeeping platform for the classification of records according to a predefined structure. Through this exercise, it was possible to gain insight into how such systems operate and to develop a realistic understanding of the results that can be expected from their implementation. The case study also made it possible to identify key requirements that archivists and records managers should consider when evaluating whether a given solution is suitable for their organizational needs.

More specifically, the exercise focused on testing commercially available solutions with respect to their ability to:

1. Automatically categorize records in accordance with organizational policies;
2. Extract metadata from records and use this metadata for their description;
3. Identify personal information within defined records series.

The test was conducted using RecordPoint’s Records365 platform. The records used to evaluate the system were drawn from the NATO Archives.

1. Identification and Preparation of the Dataset

The definition of the dataset represents a critical stage in any AI-driven project, as data must be both available and of sufficient quality. In the context of archival and records management, data can be understood as individual records or aggregations of records. Consequently, the definition and preparation of the dataset constitute a phase in which archivists and records managers can play a key role, given their knowledge of both the limitations and the potential of the datasets involved, or, in recordkeeping terms, of records aggregations.

The way in which a dataset is identified and prepared directly affects the robustness of the results. Documenting dataset biases and limitations therefore also entails documenting the parameters that inform decision-making, making this a project phase in which particular care and effort are required. Prior to the initiation of any AI project, archivists should identify a suitable dataset and ensure an appropriate level of control over it. This includes the availability of relevant metadata and the presence of well-defined structures.

In this context, archivists should assess the biases and potential of groups of digital records regarding the following aspects:

- **Availability**: Is a suitable dataset available for the intended project?
- **Quality and quantity**: Are the records of adequate digital quality (e.g. scan quality, OCR availability, and methods used)? What is the temporal coverage, size, and overall volume of the dataset? Is the number of records sufficient for the intended purpose? Are metadata consistently associated with the records?
- **Structure**: Are the records structured, and if so, how? Are filing plans, classification schemes, or ontologies available for the records? Is this information available in a structured or tabular format, and can it be shared with external stakeholders?
- **Security and privacy concerns**: Can the dataset be shared with external stakeholders? Are there privacy or confidentiality issues? Do the records contain personally identifiable information (PII)?
- **Transportability**: Can the dataset be easily transferred to the AI platform, and through which means? Is the use of cloud-based solutions viable?

Regarding data quality, one significant limitation concerns the use of historical datasets. Archivists are aware that two documents belonging to the same records series but created decades apart may exhibit both similarities and substantial differences. Such variations can pose challenges for AI models, which may fail to recognize underlying similarities and instead emphasize surface-level differences. As is often noted, the performance of an AI model is directly dependent on the quality of the data provided: the lower the level of noise in the dataset, the more reliable the model's output. These issues have a direct impact on the data training phase.

Before being used, data are typically transformed by data analysts. At this stage, the absence or inconsistency of metadata in the original dataset may become particularly problematic, further affecting data quality. This reinforces the need for standardized and well-structured metadata to be associated with datasets from the outset.

Finally, issues related to information security, privacy, and data transfer must also be considered, as datasets often need to be shared across organizational and technical boundaries. Dataset preparation constituted a significant phase of the NATO Archives–RecordPoint case study as well, albeit with certain limitations, which are discussed in the following paragraphs.

Table 4 – Logs for dataset identification and preparation

Recommended Logs for this phase	Available Logs for this phase
– Logs of any operation performed on dataset	– Dataset Full Log – Archival Series list/Archives Filing Plan – Organization Taxonomy list

PLANNING PHASE

1.1 Information about the purposes and goals to be achieved by getting the datasets processed by the AI model

The scope of the project is to assess existing AI technologies and their ability to address the challenges posed by non-aggregated, unarranged, or decontextualized records in both the current and semi-current phases of the records lifecycle. The project focuses on testing the metadata enhancement

capabilities of Records365, an electronic document and records management system (EDRMS) developed by RecordPoint. The dataset used for the project consists of several series of publicly disclosed records from the NATO Archives.

The CU05 team defined a set of functional requirements for the application. Given a collection of digital items (digitized documents) lacking metadata, the application is expected to identify relevant metadata elements and populate a structured output (e.g. a spreadsheet) accordingly. Specifically, the application should be able to:

1. Aggregate digitized documents into clusters based on criteria such as creators, topics, objects, identifiers, or signatories;
2. Extract metadata from digitized documents and generate structured metadata lists;
3. Rename documents in accordance with predefined naming conventions;
4. Flag items for which some or all metadata elements cannot be identified or captured;
5. Capture additional metadata elements, such as signatories, the original security classification of the document, public disclosure notices, and agenda items;
6. Identify and flag items that do not constitute NATO documents (e.g. national documents);
7. Perform text summarization on selected items and/or series of documents;
8. Propose semantic tagging of the collection based on controlled vocabularies and ontologies, including entities such as events, places, and people.

The application is expected to perform the above tasks (points 1–8) on two distinct sets of documents:

1. A homogeneous set of documents, such as a records series generated by a single originating office (e.g. Committee Documents), organized in chronological order and identified by a progressive numbering system;
2. A heterogeneous set of documents, such as an aggregation of records originating from multiple creators and covering different topics.

1.1.1 Information about the agents (e.g. organizations, other IT systems, people, public authority) that have made available the datasets

The dataset was provided by the NATO Archives. While NATO was established in 1949, NATO Archives was founded in 1999 and is based in Brussels. Its primary mandate is the preservation and public disclosure of Alliance records that are older than 30 years. The NATO Archives maintains documents produced by the North Atlantic Council and its sub-committees, the Military Committee and its working groups. It also holds records originating from NATO commands and missions.

Like many archives, the NATO Archives holds a large volume of digitized and born-digital records. Many born-digital records originate from shared drives characterized by deeply nested folder structures and a lack of control over their contents. Digitized records come from various sources, and metadata accompanying individual documents are often missing or inconsistent. Among its holdings, the NATO Archives includes records of a hybrid authorship nature, such as national records containing NATO-related information, for example, a letter from an ambassador to the Secretary General. This hybridity introduces additional recordkeeping challenges, particularly regarding information access and public disclosure.

Moreover, the NATO Archives adheres to strict information security policies, with security classification applied at the item level. In some cases, different paragraphs within the same document may carry distinct security classification tags.

1.1.2 Information about the selection and preparation of the datasets

For the case studies, the NATO Archives provided a large collection of declassified and publicly disclosed Committee records spanning from the 1950s to the 1990s. These records were produced by the North Atlantic Council and its Sub-Committees, as well as the Military Committee and its Working Groups. Committees and Working Groups were established to address specific topics before reporting back to the North Atlantic Council. The fonds and series structure reflects the Committee organization and is arranged according to the reporting hierarchy within the organization, as well as chronologically. Records are identified using an alphanumeric reference system.

The NATO Archives shared with RecordPoint publicly disclosed documents belonging to 38 Committee and Sub-Committee series.

- North Atlantic Council Records - C-R
- North Atlantic Council Documents - C-D
- North Atlantic Council Verbatim Records - C-VR
- Defence Planning Committee - DPC
- Defence Review Committee - DRC
- Defence Planning Questionnaire - DPQ
- Political Committee - AC 119
- Science Committee - AC 137
- Infrastructure Committee - AC 4
- Naval Armaments Group - AC 141
- Euro Inland Surface Transport - AC 15
- Civil Defence Committee - AC 23
- Wartime Commodity Probs - AC 25
- Security Committee - AC 35
- Armaments Committee - AC 74
- European Airspace Coordination Committee - AC 92
- Protection Technical Information Committee - AC 94
- Senior Civil Emergency Planning Committee - AC 98
- Civil Aviation Planning Committee - AC 107
- NATO Pipeline Committee - AC 112
- Political Committee - AC 119
- Central Europe Pipeline Policy - AC 120
- Civil Communications Planning - 121
- Science Committee - AC 137
- Naval Armaments Group - AC 141
- Industrial Planning Committee - AC 143
- Financing Von Karman Institute - AC 168
- HQ Administrative and Security Committee - AC 184
- Air Force Armaments Committee - AC 224
- Defence Research Directors Committee - AC 243
- Planning Board for Ocean Shipping - AC 271
- Civilian Budget Committee – BC
- MBFR Ad Hoc Group in Vienna – AGV
- Study Group on the Codification and Normalization – AC 135
- NATO Army Armaments Group – AC 225
- Group of Experts on the Safety Aspects of Transportation and Storage of Military Ammunitions and Explosives – AC 258
- Conference of National Armaments – AC 259

It should be noted that the Committee names are here associated with their respective series codes. These codes serve as the sole identifiers used by the RecordPoint platform (e.g., C-R, DPQ, AC 119).

Figure 1 Fonds and series structure and reference codes

The dataset consists of textual documents in PDF format. All provided documents are digital scans of the original paper records or microfilms and have undergone optical character recognition (OCR). For most records, both English and French versions are available and are included in the dataset. Metadata associated with each digitized record are recorded in the NATO Archives Public Disclosure Register.

The Register contains the following metadata at the item level:

- Record: Progressive number assigned to each document
- Originator/Provenance: creator's acronyms or alphanumeric combination
- Document Reference/Identifier: series identifiers followed by the document year and a progressive number
- Subject Title
- Date: date of the document
- Classification: intended as the current security classification after the public disclosure process
- Language: English and/or French
- Document Control ID
- File Name

The NATO Archives provided RecordPoint with an initial dataset of approximately 9,970 records belonging to the 38 Committee and Sub-Committee series described above. RecordPoint is designed to interact with various content sources that provide programmatic access to data, including OneDrive for

Business. The records were uploaded to a shared OneDrive folder in multiple zipped packages, and access to the folder was granted to RecordPoint.

When selecting documents within each series, records from different years—spanning 1955 to 2021—were chosen, where possible, to ensure representation of both older and more recent documents.

1.1.4 Information on the constraints (deriving from copyright, privacy, contractual, other legal, technology issues, etc.) bearing on the datasets.

One of the constraints affecting the dataset used in this study derives from the information security policies of the NATO Archives and the public disclosure procedures that shape the nature of the records. In particular, the Archives' information security rules prevented the sharing of available taxonomies and filing plans with RecordPoint. Only a very limited portion of the filing plan could be provided, consisting of a flat list of the series included in the dataset along with their basic hierarchical relationships. The absence of a comprehensive, publicly disclosed filing plan restricted the ability to test the Records365 platform on a broader scale.

Furthermore, although the Archives could provide a current public taxonomy, this was limited and did not include historical taxonomies or filing plans. Historical records require accompanying historical classification structures, which change over time, and the Archives was unable to supply this supporting documentation.

While information security represented a significant barrier in this case study, it is important to note that traditional information management toolkits remain essential for the implementation of AI solutions for records classification. According to the findings of this study, records (and data) require structures such as filing plans, ontologies, and reliable metadata in order to be effectively processed by AI-based rules and classification tools.

A further constraint was technical in nature. The NATO Archives does not maintain a platform for the bulk sharing of large document collections with external entities, and data transfer involved a lengthy, manual process. As a result, only approximately 14,000 records could be shared with RecordPoint out of a potential total of around 300,000. The manual creation of zipped packages, limited to a maximum of 50 records each, required substantial time for dataset preparation. This highlights that tools enabling agile sharing of large datasets are a prerequisite for projects of this kind. The dataset used in this study did not contain records subject to copyright, privacy, or contractual restrictions.

EXECUTING PHASE

1.2.1 Information about the dataset to be processed by the AI model (structure, type of data of the dataset, history of the dataset)

The case study dataset includes only publicly disclosed records, which represent a small fraction of the total holdings of the NATO Archives. This limitation has several implications for the study. The first consequence is that only a limited number of documents from each series could be used. In fact, within a series of records, not all records are necessarily publicly disclosed. Secondly, the arrangement of the publicly disclosed records derives from the disclosure administrative process and not from their 'original order' within the series, as explained in the following paragraph.

The dataset includes records resulting from two administrative procedures for the public disclosure of NATO records. The first procedure is called the Systematic Public Disclosure Process (1), while the second is called the Ad Hoc Process (2). The two administrative procedures inform the current structure

(or arrangement) of the public records. In fact, the (1) Systematic process declassifies and makes public chronological series of Committee records according to their years and original security classification, while the (2) Ad Hoc process makes public records that pertain to a specific subject or topic, originating from different dossiers and originators, including divisions, military commands, and operations.

In addition, an Ad Hoc request (2) may include records produced by NATO nations that contain NATO-owned information. This portion of the dataset raises the question of distinguishing national from international records. It should be noted that the NATO Archives does not maintain filing plans for national records.

CLOSING PHASE

1.3.1 Outcome of the Quality Checks performed on the Dataset to be Processed by the AI Model and 1.3.2 Compliance statements (e.g. statements where the fact that the datasets are deemed to be fit for the purpose is attested to)

In order to perform automatic classification, the RecordPoint platform requires at least 50 instances of any given category of records (e.g., a record series) to train its AI model. For this reason, the dataset resulting from the Ad Hoc Process was deemed unsuitable for the model training phase, as the Archives could not provide the minimum number of records related to the same topic or originating from the same author. Consequently, it was also not possible to conduct testing related to the identification of national records (produced by national embassies or bodies in the context of international missions, for example) versus international (NATO-only) records, as the NATO Archives does not hold a sufficient number of publicly disclosed records of these types.

As a result, the case study was limited to testing on a homogeneous set of documents only, consisting of various series of NATO Committee Documents produced under the Systematic Public Disclosure Process described above.

Further limitations were observed in relation to metadata consistency. Over the 25 years of the NATO Archives' public disclosure activities, different metadata standards have been applied to publicly disclosed records. For instance, public taxonomies have not always been consistently populated, affecting the suitability of sections of the dataset for AI testing.

Additional challenges arose from records digitized from microfilms. When scan quality was poor, OCR accuracy was compromised. Moreover, since the case study included records dating from the 1950s to 2021, certain metadata elements, such as identifiers, have undergone minor changes in spelling over time. These variations could introduce confusion for AI models, which require precise and standardized inputs and outputs to perform classification and metadata extraction tasks effectively.

In conclusion, while the case study was limited in achieving some of its initial objectives, it highlighted the importance of internal housekeeping practices within the Archives. In particular, inconsistencies and heterogeneity in metadata application can significantly hinder the initiation and success of AI-driven records management projects.

2. Produce the AI Model

Although archivists and records managers have limited control over system operations, it is important that they understand, at a conceptual level, how the platform functions. Records365 is an in-place records management system capable of ingesting data from any digital content source. The platform classifies records to determine their retention period and automatically disposes of them when disposal is due.

Automated classification is implemented in two forms: rules-based (Expert System) classification, which relies on record metadata, and machine learning (ML) classification, which is based on the content of records.

Expert Systems are designed to solve complex problems by reasoning through structured bodies of knowledge, primarily represented as if–then rules. A rules-based system operates via pattern recognition, requiring a codified set of rules to function. In archival terms, these rules can correspond to classification codes; for example, if a document belongs to a specific class, the automatic classifier applies the appropriate actions. The advantages of Expert Systems include reliability—responses are generated solely based on codified rules, eliminating the risk of hallucination—and accountability, as the system provides a clear description of how decisions are made. These features are particularly relevant in the information management domain. Among the disadvantages are potential computational complexity for simple tasks and the requirement for clear, well-defined rules applied at the item level.

Machine learning operates differently: given a sufficiently large dataset, the system can infer patterns and logic from the data. RecordPoint's Classification Intelligence feature uses ML to categorize records consistently according to a defined classification taxonomy. The software classifies records based on their textual content and enriches them by extracting personal identifiers, named entities, and other signals from both text and metadata.

As noted above, a minimum of 50 sample documents pre-classified according to the relevant ontologies is required for automatic classification using ML. No additional supporting documentation is needed to train the system, provided that records have been pre-classified. The platform employs a supervised learning strategy, with a human-in-the-loop who can confirm or reject ML outputs. This allows non-technical users to review the training documents and triage them as necessary.

Regarding metadata enhancement, the system extracts extrinsic metadata from external content sources and intrinsic metadata from within record files. Extrinsic metadata, defined by external systems, are incorporated using pattern-matching rules, while intrinsic metadata are derived by AI technologies from the record text. In the latter case, a "signaling" process calculates new metadata based on the harvested or mined information. This process includes Named Entity Recognition, allowing the identification of organizations or individuals referenced in the record text. Record text is extracted through optical character recognition (OCR), though the quality of the resulting text may vary, and OCR processing can be computationally expensive. The Intelligence Signaling feature scans all data processed by the RecordPoint intelligence engine to detect privacy-sensitive information, including Personally Identifiable Information (PII) and Payment Card Industry (PCI) data.

The platform relies on Microsoft's Presidio SDK for the detection of PII, PCI, and related signals (https://microsoft.github.io/presidio/). Custom recognizers are not used; regular expressions (regex) are employed for specific signals, while most other signals are derived from Presidio's built-in catalogue.

Records365 also provides additional AI functionalities, including question answering, text synthesis, translation, and image recognition. While technologically straightforward, these features can be costly to implement at scale.

Finally, the RecordPoint Data Trust Platform is compliant with, or designed to support, ISO 15489, the Data Management Capability and Assessment Model (DCAM), the California Consumer Privacy Act (CCPA), and the General Data Protection Regulation (GDPR).

Table 5 – Model development logs

Recommended Logs for this phase	Available Logs for this phase
– Logs of the operations conducted to develop the AI model	

3. Train the AI Model

The training phase[14] defines what, in archival terms, may be considered the records' "context." When a model is trained, the system is taught what constitutes a record and how it should be categorized. At the same time, the parameters and thresholds for acceptable classification outcomes are established, determining whether a record is automatically associated with a class or flagged for reassessment.

Training models with historical records presents additional challenges. For example, a series of records spanning several decades may maintain the same reference code, yet small variations can occur—such as the replacement of a slash or space with a parenthesis, or the presence of both typewritten and computer-generated letters within the same series. These examples illustrate why data identification and preparation, as well as involvement of information professionals as subject-matter experts, are critical steps in AI model training.

Furthermore, AI models require regular updates. In supervised learning, a model does not adapt until it is provided with new training data. Consequently, if the operational environment changes, the model must be retrained. This necessitates careful planning for training cycles in both project timelines and budgets.

Table 6 – Training logs

Recommended Logs for this phase	Available Logs for this phase
– Logs of the training activities	– 1st Model Outputs and Performances Log – 2nd Model Outputs and Performances Log

PLANNING PHASE

3.1.1 Methodologies followed to train the AI model

The following points summarize the workflow for setting up and training the model:

- Archives must provide a filing plan or ontology of the records so that rules can be configured accordingly within the system;
- A machine learning model is built within the Records365 platform;
- Archives must provide sample documents pre-classified according to the relevant filing plan or ontologies;
- Models based on the training set are evaluated using K-Fold cross-validation;
- Archives must provide at least 50 pre-classified sample documents per category to train the model;
- Archives must provide additional records within the same categories to test the model.

For the case study, the NATO Archives provided the available taxonomy and associated records for the training phase, subject to the limitations previously described in the dataset preparation section. RecordPoint selected the series to train the model based on the number of available records in each series.

RecordPoint employs an algorithm that prioritizes recent records with a minimum and maximum amount of text data for inclusion in the training set. The training set is automatically supplemented with previously misclassified records to enable the model to learn from its errors. Models are evaluated and selected

[14] See RecordPoint website dedicated to describing the training phase for classification, https://help.recordpoint.com/hc/en-us/articles/5385817601423-Training#trained-models.

using K-Fold cross-validation[15]. In operational contexts, where new records continuously enter the platform, models can stagnate if not updated, making regular retraining essential. Each new training cycle incorporates records that were previously misclassified to improve model accuracy. Within the context of this case study, two cycles of retraining were conducted, as detailed below.

3.1.2 Design of the activities carried out to train the AI model

A machine learning model is constructed within the Records365 system, as illustrated in the screenshots below. The images display the list of series to which the records belong, according to the provided filing plan.

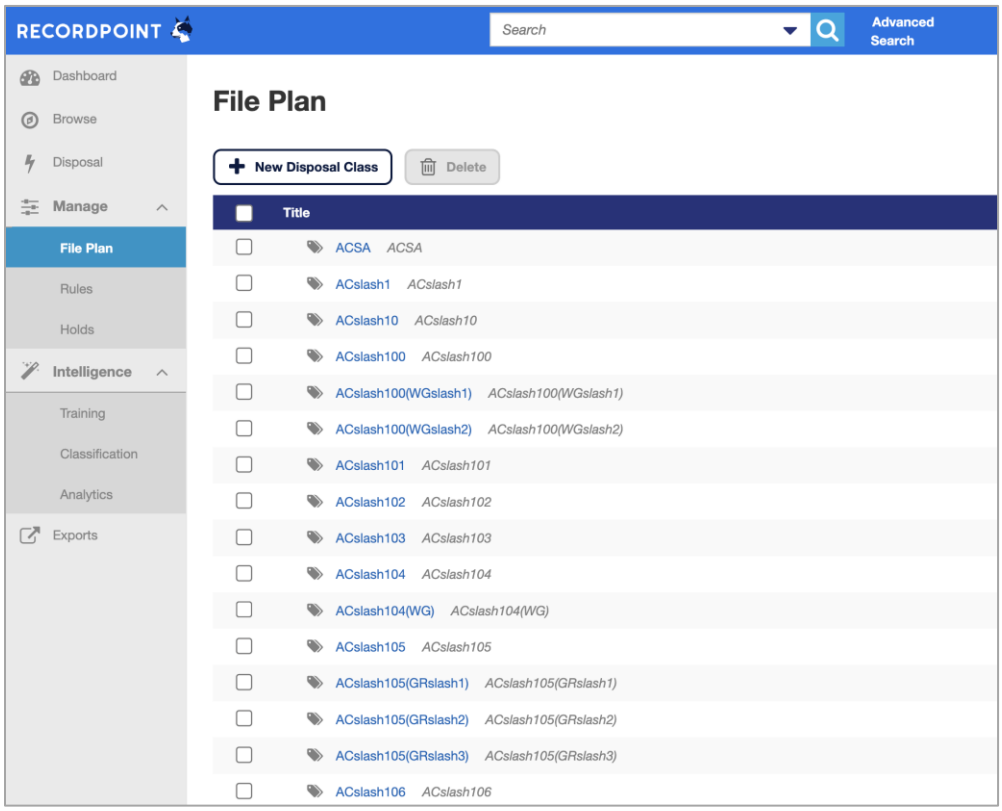

Figure 2 – Series list within the Records365 System

The following image illustrates the hierarchy among the series and the built-in workflow that defines the categorization rules.

[15] Cross-validation is a statistical method used to estimate the skill of machine learning models. See <https://machinelearningmastery.com/k-fold-cross-validation>.

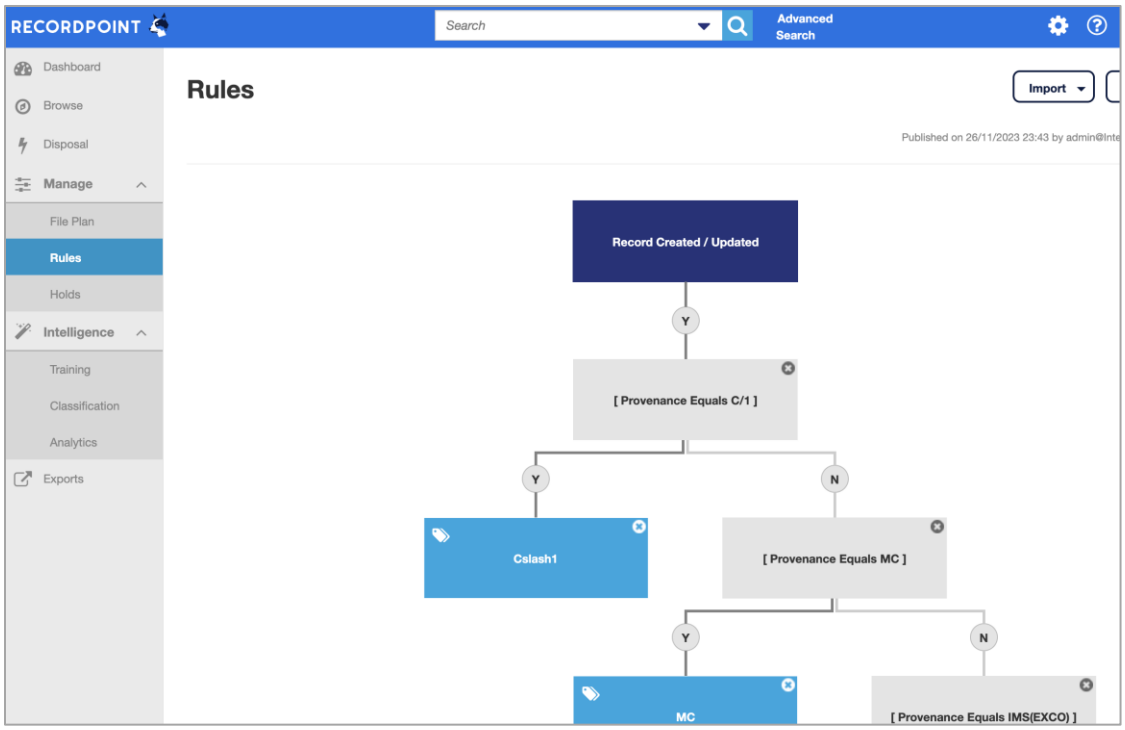

Figure 3 – Series hierarchy within the Records365 System

The model was then trained using a set of records belonging to the same series, as shown in the screenshot below. In this example, the training set corresponds to the Science Committee records (AC 137). Each record title includes the Science Committee identifier AC/137, which corresponds to the disposal class “ACslash137.”

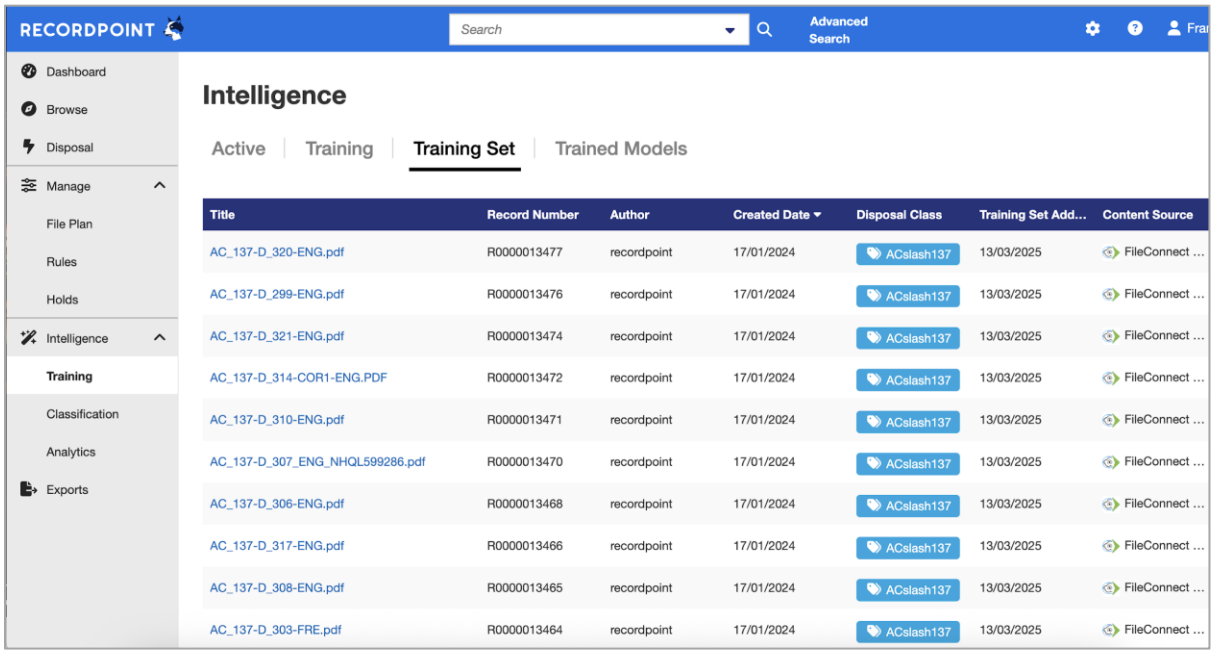

Figure 4 – Records training phase

EXECUTING PHASE

3.2.1. Information about the Training Datasets used to Train the AI Model

Among the records and series identified during the dataset preparation phase (1.1.2), only the top 18 categories were selected to train the model, primarily based on the number of available records in each category. Although additional categories could be included, training time increases as more categories are added. Selection criteria also considered recency and file size.

Each category requires a minimum of 50 records for training, although a larger number is preferable. The training set is capped at a maximum of 500 records per category. When more than 50 records are available, recent records are prioritized over older ones. Records365 employs an algorithm that selects the training set to favor recent records with a minimum and maximum amount of text data. Each training run incorporates previously misclassified records, enabling the model to learn from its errors.

Another factor in selecting records for training is the volume of text contained within each file. Very small files with limited textual content are excluded, as they provide insufficient data for model training. Files between 5 KB and 500 MB are preferred. These parameters are configurable, but the default settings have proven effective in most cases and are rarely adjusted by RecordPoint.

For the case study, records from the following series were selected for the AI model training phase:

- North Atlantic Council Records - C-R
- North Atlantic Council Documents - C-D
- North Atlantic Council Verbatim Records - C-VR
- Defence Planning Committee - DPC
- Defence Review Committee - DRC
- Defence Planning Questionnaire - DPQ
- Science Committee - AC 137
- Infrastructure Committee - AC 4
- Security Committee - AC 35
- Political Committee - AC 119
- Central Europe Pipeline Policy - AC 120
- Naval Armaments Group - AC 141
- Industrial Planning Committee - AC 143
- Conference of National Armaments – AC 259
- MBFR Ad Hoc Group in Vienna – AGV
- Study Group on the Codification and Normalization – AC 135
- NATO Army Advisory Group – AC 225
- Group of Experts on the Safety Aspects of Transportation and Storage of Military
- Group of Experts on Transportation of Military Ammunitions and Explosives – AC 258

The screenshot below provides an overview of Records365's performance following the first model classification exercise. The first column, Record Disposal Class, indicates the series assigned by the classifier—for example, ACslash119 corresponds to the Political Committee (AC 119) records. The third column displays the classification success rate (Classification Skill) for each series. Some series achieved very high classification rates (1.000), while others showed lower performance (0.706, 0.647) or insufficient accuracy. To improve the performance of series with lower success rates, additional sample records are required for retraining.

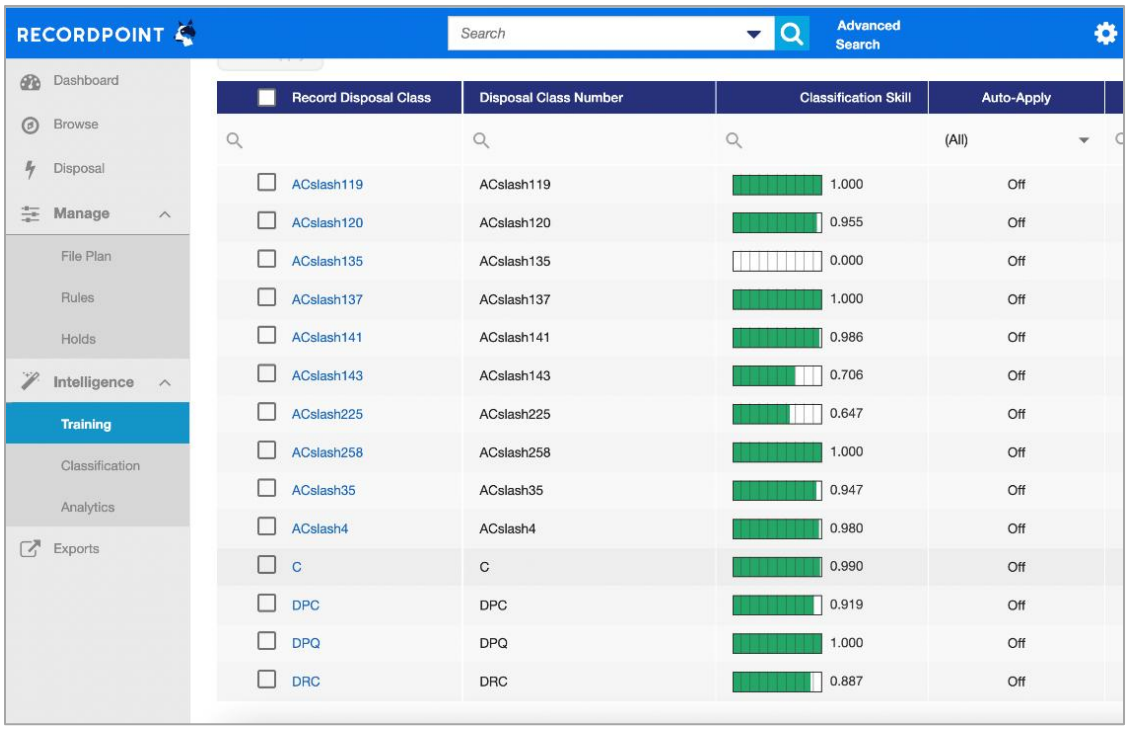

Figure 5 - Training Results for the Selected 14 Record Classes – First Model (22/04/2024)

This training exercise led to an overall improvement in classification performance following further training and the development of a second model. However, some results remain below acceptable levels.

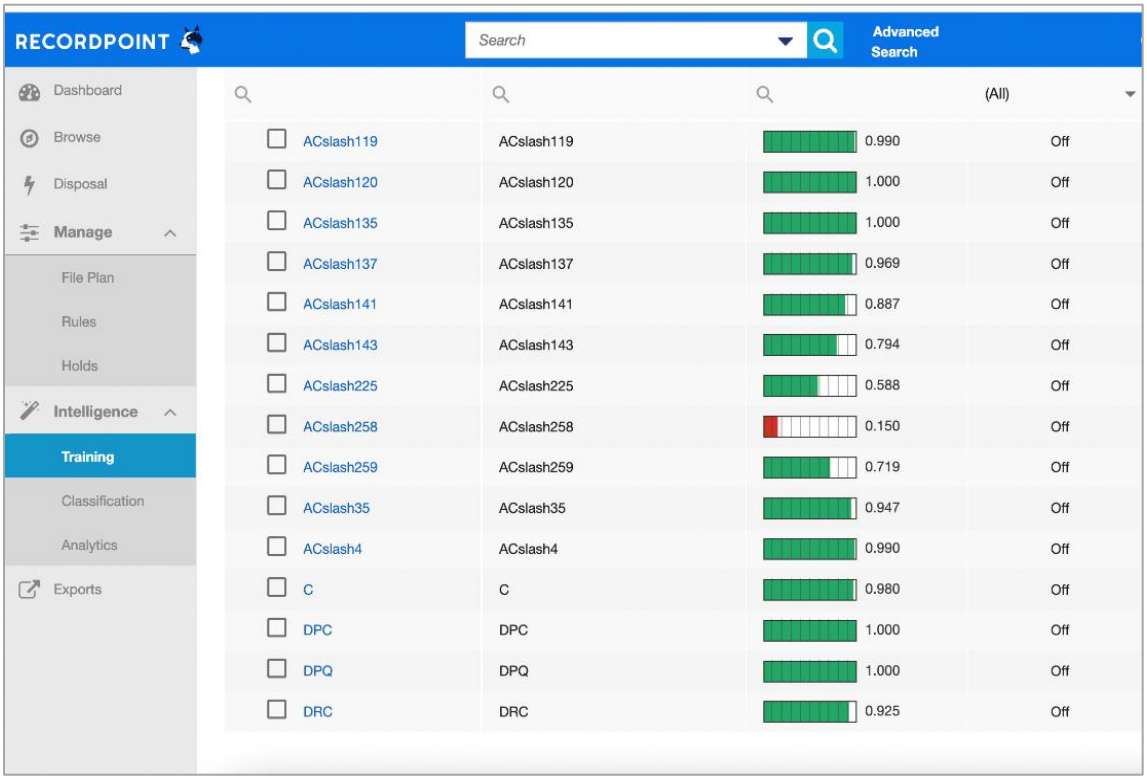

Figure 6 - Training Results for the Selected 14 Record Classes – Second Model (05/08/2024)

The screenshot below illustrates the classification tool applied to a small sample of records. Users can support model tuning by either accepting the category suggested by the machine learning model or reclassifying the record manually. Manual tuning is an integral part of the classification exercise.

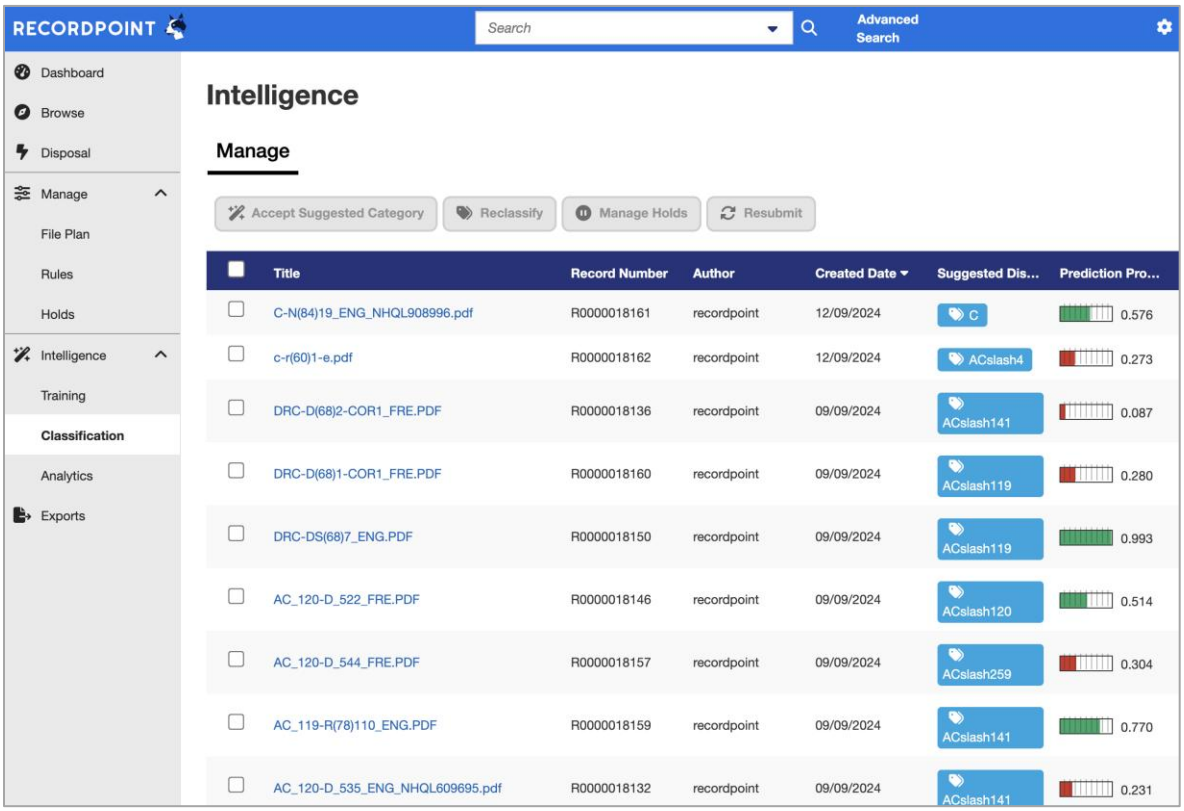

Figure 7 – Application of the classification tool

CLOSING PHASE

3.3.2 Outcomes of the quality checks performed on the training dataset

The models identified four series—AC 143, AC 225, AC 258, and AC 259—as particularly challenging to classify, as shown in the images above. Several factors may explain these lower performance ratings. First, these series span a wide range of dates and include some low-quality images.

Records within the AC 259 series, corresponding to the Conference of National Armaments, are further divided into the following sub-series:

- AC 259/D – Document
- AC 259/DS – Decision Sheet
- AC 259/N – Notice
- AC 259/WP – Working Paper

This sub-series structure is common in NATO records, where identifiers such as D, DS, N, and WP are used to classify subsets of records within a series, as observed in other cases described in this study.

The first model was trained using 50 instances of AC 259/D records dated from 1967 to 1989. It was then tested with a combination of AC 259/DS, AC 259/WP, and 20 additional AC 259/D records. Since the

model had been trained on only one sub-series (AC 259/D) but tested on multiple sub-series, its performance was relatively low. The second model incorporated two sub-series (AC 259/D and AC 259/N) during training, resulting in improved classification accuracy.

These results suggest that the more precise and complete the ontology or filing plan provided at the outset, the more accurate the classifier's performance will be.

Table 7 - Tab. AC 259 - Conference of National Armaments

Rules	Model Training	Model Testing	Score	Second Model	Score	Quality of the Scans	Date Range
AC 259	AC 259/D	AC 259/D AC 259/DS AC 259/WP	Low	AC 259/D AC 259/N	Higher and acceptable	Includes some low quality	1967-1989

In addition, when the quality of the images and the number of testing records are insufficient, the model cannot achieve acceptable performance, as observed for some AC 259 records.

A similarly low classification rate was obtained for the series Group of Experts on Transportation of Military Ammunitions and Explosives (AC 258). In this case, the Archives could provide only 20 sample documents for model training, and the quality of the scanned images was relatively poor. As a result, it was not possible to develop and test a second classification model, as the number of records was insufficient to train the model effectively.

Table 8 - Tab. AC 258 - Group of Experts on Transportation of Military Ammunitions and Explosives

Rules	Model Training	Model Testing	Score	Second Model	Score	Quality of the Scans	Date Range
AC 258	AC 258/D	AC 258/D AC258/D(ST)WP	Low	Insufficient number of records available	Low	Poor	1967-1989

A different case is represented by the NATO Army Advisory Group – AC 225 Subordinate Committee (1967–1989), which assumed the duties previously managed by the Naval Armaments Group, along with the related documents produced in the AC 141 series. The model was trained using 50 AC 225–D documents and 50 AC 225–N documents.

These records reflect, in a relatively loose manner, the shift of responsibilities between the aforementioned Groups and the Advisory Group. Manual notations were used by registry personnel at the time to identify the correct document reference. As shown in the image below, although three references were typewritten on a document, the accurate reference was the last one (AC/225-N/9). The model could not predict this outcome unless cross-referencing rules were explicitly incorporated. In such cases, the model struggles because it cannot make decisions outside the predefined set of rules.

This example further illustrates that institutions intending to implement AI-based classification projects must ensure that filing plans, retention and disposition schedules, and taxonomies are available in machine-readable form and accurately reflect historical and organizational changes over time.

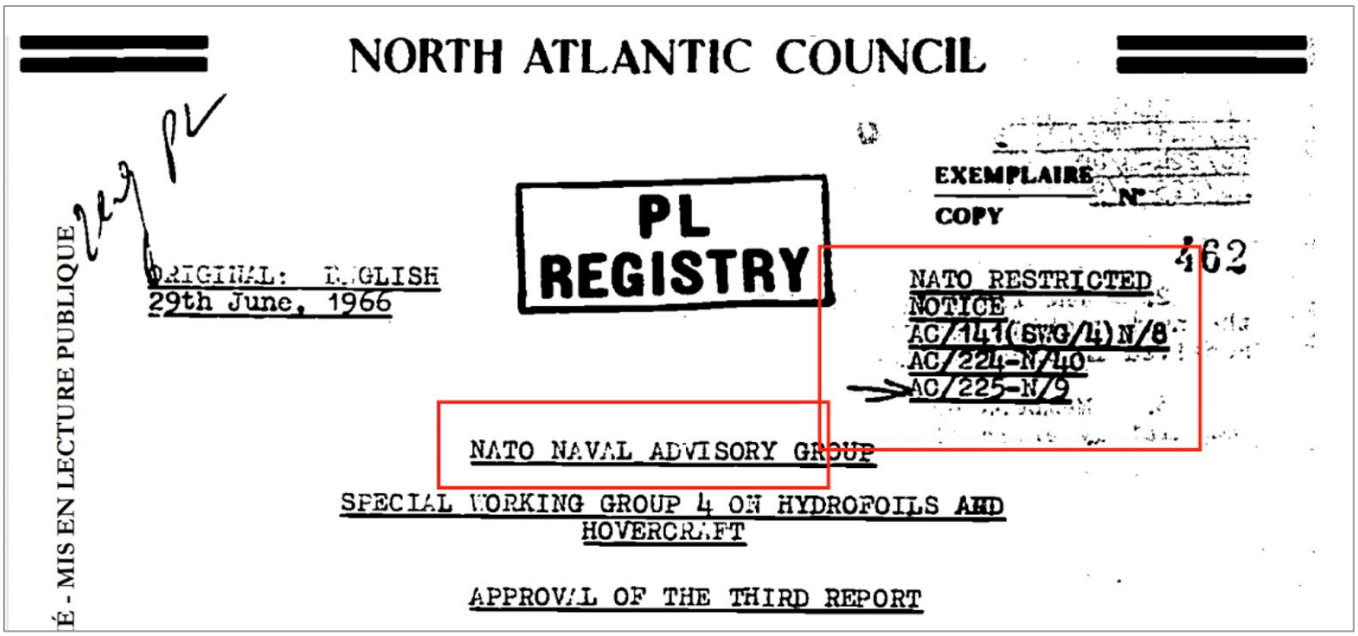

Figure 8 – Example of cross-referencing

Table 9 - Tab. AC 225 - Army Advisory Group

Rules	Model Training	Model Testing	Score	Second Model	Score	Quality of the Scans	Date Range
AC 225	AC 225/D AC 225/N	AC 225/D AC 225/N	Low	1967-1989	Higher and acceptable	Fine but cross references with other series pose challenges	1967-1989

A final remark concerns the language of the records. RecordPoint currently supports English for machine learning-based classification. Within the NATO Archives corpus, 4,492 records were in French and 5,483 in English. Although the ML system was configured only for English, testing demonstrated that it also performed effectively with French records.

When classifying French documents, the system analyzes the text content and automatically assigns categories based on the information contained within the document. This approach is effective because the French records constitute approximately half of the corpus. The text pre-processing assumes English as the primary language; however, this does not prevent the model from incorporating terms from other languages during classification.

4. Evaluate AI Model

Table 10 – Model evaluation logs

Recommended Logs for this phase	Available Logs for this phase
– Logs of the measurements and controls carried out to evaluate the AI model	– Models Performances Report

PLANNING PHASE

Records365 target metrics are defined by users according to the specific context and objectives of a project. Common metrics include the completeness of record ingestion (typically targeting 100%), processing speed, and classification coverage. Machine learning models are evaluated based on test accuracy or F1 score, depending on the distribution of categories. Typical expectations are approximately 80% accuracy or an F1 score of 0.80.

For each ML-classified record, the platform provides a **probability score** indicating the confidence level of the system in its suggested classification. Misclassified records identified by users are fed back into the ML model to mitigate potential biases.

It should be noted that ML classification relies exclusively on linguistic information contained in the records; numerical data are generally discarded. Records dominated by numerical content are likely to produce low confidence scores and require manual classification by a human user.

EXECUTING PHASE

Following the second round of training, the classification exercise yielded positive results for several record series that were correctly classified. As previously discussed, certain series continued to present challenges for the classification system.

In contrast, greater difficulties were observed in relation to additional metadata creation and the identification of personally identifiable information (PII) within the dataset. Specifically, the metadata enhancement feature includes a binary field, “hasPersons”, which returns either “yes” or “no.” While this feature may be useful in certain contexts, it does not fully satisfy the archivist’s need to extract from the record text the names of persons mentioned in the document or the author of a letter or memo. Projects with such requirements may benefit from a different approach, potentially involving tools with more advanced optical character recognition (OCR) capabilities.

The platform’s PII detection feature, which flags the presence of PII in documents, produced several false positives, as evidenced in the Metadata Enhancement log. This feature relies on Microsoft Presidio, and calibration improvements would be necessary to reduce false positives. It is important to note that NATO Archives records were not used to train the Microsoft Presidio model; therefore, the system had not been adapted to this specific dataset.

The screenshot below illustrates one of the false positive results from the Metadata Enhancement exercise, indicating Has PII – Personal Identifiable Information/Has Person. The full record, C-N(84)19_ENG, is provided in the appendices.

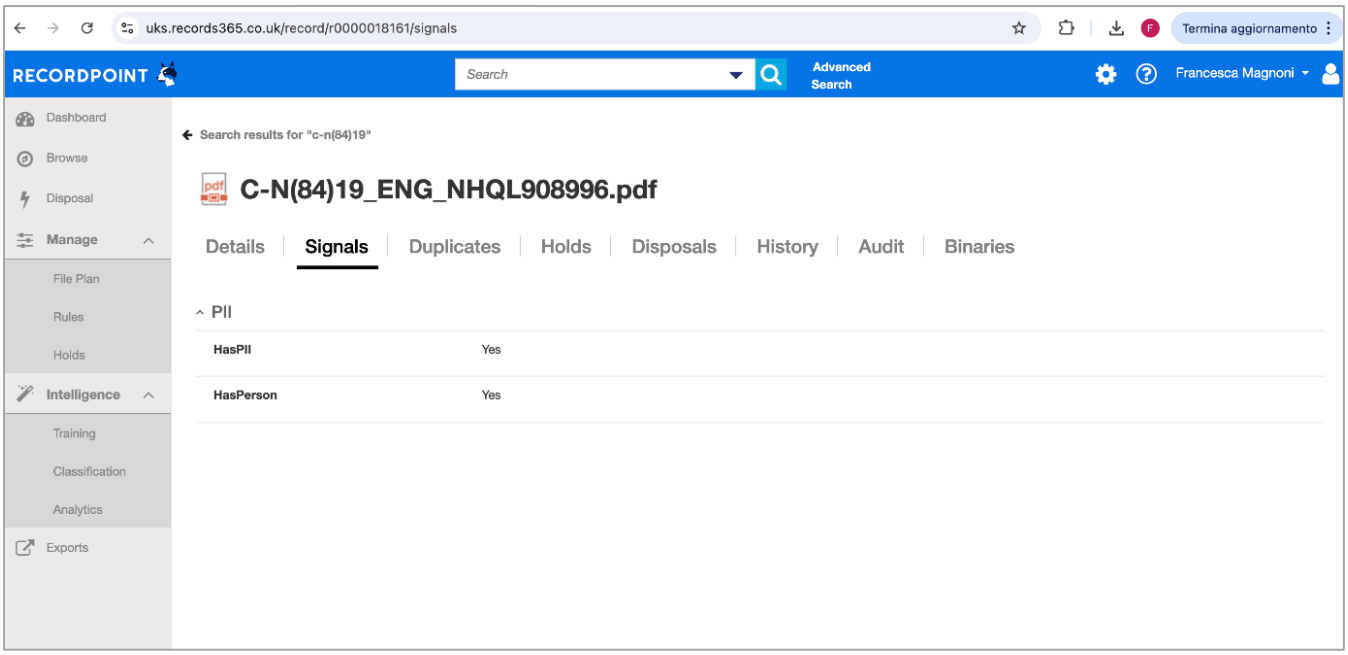

Figure 9 – Results from the Metadata Enhancement exercise

The final screenshot illustrates an example of text summarization. The summarized document is a sixteen-page record of the North Atlantic Council with the identifier C-R(60)1, which is also provided in the appendices. The proposed summary is very high-level, highlighting that when archives plan AI-assisted description projects, the limitations of automated summarization should be carefully considered.

Due to the computational costs associated with this type of exercise, only a small number of records were processed.

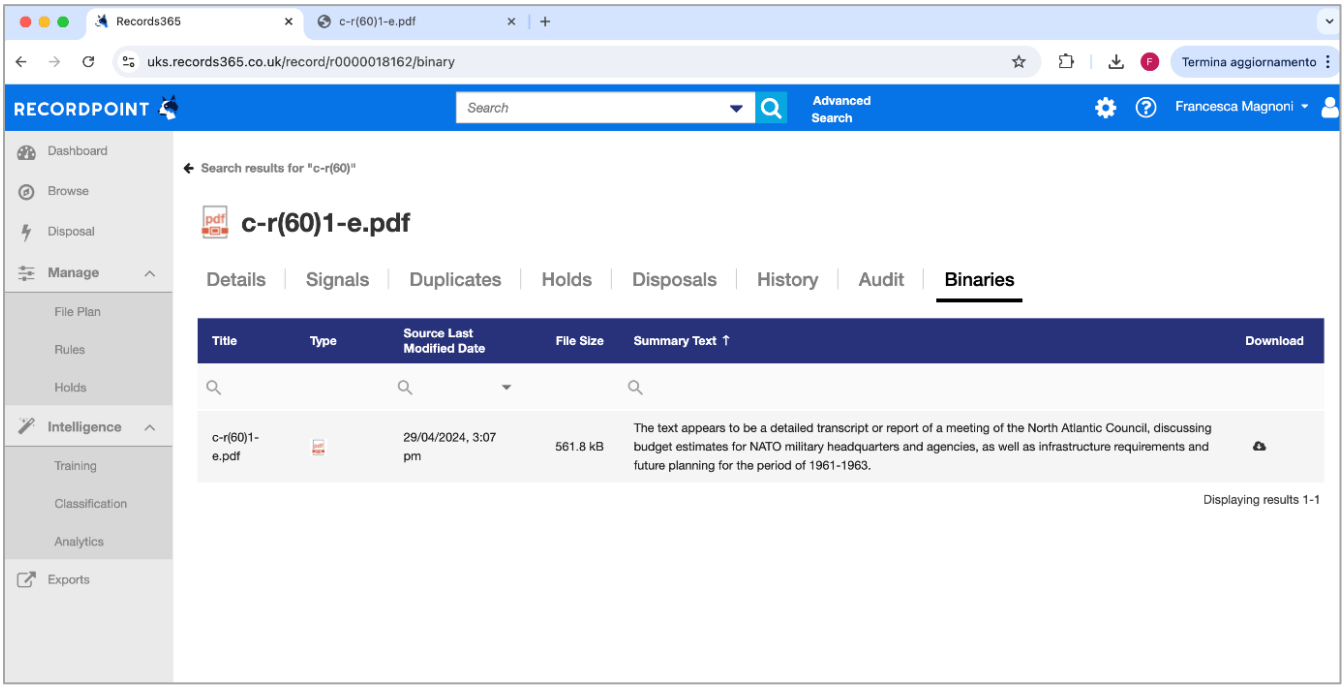

Figure 10 – Example of Text Summarization for NATO Council Record C-R(60)1

Table 11 – Tab. Text Summarization Workflow as of July 2024

Text Summarization	Proof of concept level	Example	Comments
	If the text is less than 100 characters, it summarises it in 1 paragraph. If it is longer, it goes to ChatGPT that makes the summary, and if the text is in French, first it translates it and then it summarises	C-R(60)1 English	Archives should keep this in mind if considering text summarization to support archival descriptions. Limitations are also evident on the output itself and more detailed instructions on the outputs could be provided - i.e. identify the author, the action, the object.

5. Implement AI Model

Table 12 – AI model implementation logs

Recommended Logs for this phase	Available Logs for this phase
– Logs of the actions and decisions made by the AI model – Logs of any event that might have impacted on the AI model – Logs of the interactions between the AI model and other IT systems	

6. Improve AI Model

Table 13 – Improving AI model logs

Recommended Logs for this phase	Available Logs for this phase
– - Logs of the modification activities	

7. Monitor Operations

Table 14 – Monitoring operations logs

Recommended Logs for this phase	Available Logs for this phase
– Logs of the monitoring operations – Logs of the tools used to carry out monitoring operations – Logs of other systems which work dependently on the AI model	

Table 15 provides an overview of the project documentation available. It summarizes the materials collected and organized to support the case study, including initial project proposals, records logs, machine learning reports, and other relevant documentation used to track and evaluate the AI-enhanced recordkeeping exercise.

Table 15 – Overview of the available project documentation

	AI Application Lifecycle	**Project Documentation**
InterPares Team/NATO Archives Documentation	1. Identification and preparation of datasets	InterPares Case Study Report
		Dataset Full Log: NATO Archives Public Disclosure Register
		Archival Series list/Archives Filing Plan (if available)
		Organization Taxonomy list (if available)
		Organization Acronyms list (if used)
		Organization's Security Classification list (if used)
		Relevant Organizational Policies and AI Policy (if available)
RecordPoint Documentation	2. Produce AI model	AI Platform high-level description (from the survey)
		RecordPoint White Paper
	3. Train AI model	Model Training Records Log
		Model Training Records Log by Series
		Model Training Records Log by Language
		1st Model Outputs and Performances Log
		2nd Model Outputs and Performances Log
	4. Evaluate AI model	Models Performances Report
		Metadata Enhancement Log
	5. Implement AI model	Models Performances Report
	6. Improve AI model	Models Performances Report
	7. Monitor operations	InterPares Case Study Report

Table 16 – Simplified Version of the Paradata Table and Available Resources with Colour Coding

	Phases of Data Collection		
	1. Planning	2. Executing	3. Closing
1. Identification and preparation of datasets			
2. Produce AI model			
3. Train AI model			
4. Evaluate AI model			
5. Implement AI model			
6. Improve AI model			
7. Monitor operations			

8. Conclusion

This case study demonstrates the essential role of archivists in AI projects for recordkeeping. Archivists are crucial in documenting processes, identifying relevant data logs, and describing datasets, including their limitations and characteristics. While collaboration with data scientists and IT specialists is required for technical phases such as model production and implementation, archival expertise ensures transparency, accountability, and meaningful use of AI systems. The study highlights that effective AI in archives depends not only on

12.3 Case study of Regione Emilia-Romagna . Artificial intelligence in document classification: classifying through unsupervised actions a series of public records

1. Working Group

- **Regione Emilia-Romagna**: Riccardo Righi, Barbara Sorace, Simone Sorce, Marianna Tascone, Patrizia Varini
- **Università di Bologna**: Paolo Torroni, Elena Palmieri
- **Interpares**: CU05 Study “The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas”

2. Researchers

- **Archival experts**: Riccardo Righi, Barbara Sorace, Simone Sorce, Marianna Tascone,
- **AI experts**: Paolo Torroni, Elena Palmieri, Patrizia Varini
- **Interpares**: CU05 Study “The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas”

3. The case study

The Polo archivistico dell’Emilia-Romagna - ParER (Emilia-Romagna Digital Archival Centre) is a trusted digital repository of the Regione Emilia-Romagna (Italy), responsible for the long-term digital preservation of digital records transferred from all regional public administrations. It functions as a centralized archive to which participating institutions contribute their own archives, benefiting from a high-level professional, archival, technological and organisational service.

It is considered one of the Italian best practices in preservation of digital archives, and its mission is to develop policies and procedures for record-making, record-keeping and record-preservation, in order to ensure that all data transferred from the public administrations can be safely stored, in order to be accessible in the forthcoming years.

ParER preserves different digital record typologies: administrative, educational, cultural, health care, etc. Up to now, it has preserved over 3 billion records, coming from over 100 public administrations.

ParER has its own technological infrastructure, with two physical datacenters (the main one in Bologna and the backup one in Parma). Its services are guaranteed by a totally own preservation software called “Sacer”, the design, improvement and development of which are under the total control of ParER staff, which amounts to over 20 persons and 15 external professional support.

ParER is also involved in the definition of national models and rules on recordkeeping and digital preservation and participated in various international projects: InterPARES, as partner of the project from 2010, and E-ARK Project as member of archival advisory board.

The study investigated the possibility of automating the assignment of classification classes to administrative documents. The Regione Emilia-Romagna (hereafter RER) provided approximately 3,000 administrative documents (executive acts and records) for the case study. All documents are original digital files in PDF format.

Given the lack of pre-classified documents required to train supervised systems, the research focused on the feasibility of unsupervised classification techniques. The effectiveness of these methods was subsequently validated by comparing the results with previously correctly classified samples.

The project addressed several complexities intrinsic to the administrative documentation:

- text heterogeneity: long documents lacking standardized structures.
- complexity of the classification plan: a system based on distinct classes of meaning related to different functional areas.
- scarcity of samples: limited availability of homogeneous documents for testing.

The research followed two parallel tracks:

- academic experimentation with University of Bologna.
- internal experimentation at ParER.

The following table describes the main characteristics and differences between the two experiments.

Table 17 – Main features of the two experiments

Features	**Academic Experimentation with Università di Bologna**	**Internal experimentation at ParER**
Dataset	A selection of 1.728 managerial acts (*provvedimenti amministrativi,* adopted to conclude complex processes), characterized by long texts and heterogeneous forms.	A selection of 1.519 registry records (*documenti di protocollo,* Incoming and Outgoing Correspondence).
Nature of the Data	Public records already available online. For the purposes of the study, a copy was stored in a shared storage space provided by the University of Bologna.	Public records not available online. The records were exported from the organization's document management system and stored in a protected corporate SharePoint folder.
Objective	To automate the classification of documents according to the National Classification's Plan Model for Regions.	To automate the classification of documents according to the official classification plan (*titolario*) of the RER.
Approach	Unsupervised classification was tested using models such as BART, T5, RoBERTa (encoder/decoder and masked language models) and GPT-4o, Gemini, and Llama (Large Language Models – LLMs).	This phase employed Mistral Small (open source) and GPT-4.1(proprietary) models, evaluating their accuracy within a real-world production context.
Duration	2023-2025	2025

3.1. Methodology and Operational Phases

In both experiments, the research was structured as follows:

- text preprocessing: identification of the most relevant document sections for classification and the merging of multiple files relating to a single record;
- model experimentation: evaluation of the degree of accuracy and the resources (time/costs) required by each model;
- analysis of results: comparison between the performance of the models.

3.2. Academic experimentation with the Università di Bologna

In 2023, RER initiated a collaboration with the Università di Bologna to support the research activities of Elena Palmieri, a doctoral student from the Department of Computer Science – Science and Engineering. The subject of her doctoral research "Scarce Data and Complex Texts: the Role of LLMs in Automatic Text Classification", which concluded in 2025, focused on the potential use of artificial intelligence models for the unsupervised classification of administrative documents produced by RER. The results were presented on 09/04/2025 by Professor Paolo Torroni and Elena Palmieri during a webinar organized by Formez PA[16], an institute that supports the implementation of reform and modernization policies for the Italian Public Administration through training activities, recruitment of public personnel, support and technical assistance also for small municipalities, as well as the promotion of innovation and the strengthening of administrative capacity.

3.2.1. Description of Data and Domain

For the experiment, RER selected 1.728 public records that were already available online. For the purposes of the study, a copy was stored in a shared storage space provided by the Università di Bologna. All documents are in Italian and were originally in PDF format. For the experiment, managerial acts were considered, that is, decisions made on various topics which were to be linked to the classification plan (Titolario) model of the Regional Councils (in Italian: "Titolario delle Giunte regionali" hereafter TGR), developed by a working group established in 2002 by the Ministry for Cultural Heritage and Activities. This is a specific Titolario for the Regions, which organizes documents according to regional functions and competences (e.g. environment, territory, energy, health), in accordance with current legislation regarding the creation, management and preservation of documents. The TGR contains 19 first-level classes, each of which has a brief description in natural language (around 600 characters) that can be used as external knowledge for the classifier. The hierarchy has two levels of depth, but the experimentation focused on the first level.

TITOLO 1

ORGANI DI GOVERNO, AFFARI GIURIDICO-ISTITUZIONALI E COMUNICAZIONE

Comprende le funzioni e le attività relative alla configurazione degli organi di governo e amministrativi della Giunta (presidenza, vicepresidenza, giunta, direzioni centrali ecc) in materia di indirizzi organizzativi e di funzionamento generali, di rapporti istituzionali con organi statali, regioni, enti locali, comunità montane, enti strumentali, enti vigilati, partecipate regionali, istituti pubblici e privati, ecc., rapporti con organismi extranazionali e relazioni internazionali; organismi generali e intersettoriali; comunicazione; rappresentanza; affari legali.

Figure 11 – Example of a description of a TGR class

Characteristics of the Gold Dataset

The Titolario of the Regione Emilia-Romagna (hereafter TRER) has evolved over time, and the documents classified according to it do not have direct correspondences with the classes of the TGR, making it impossible to use TRER classification as an evaluation tool. Therefore, it was deemed useful

[16] L'intelligenza artificiale nella classificazione dei documenti: potenzialità e limiti, <https://eventipa.formez.it/eventi/4071d1d7-89aa-4bc6-876c-19b79b29d2cd>

to prepare a Gold Dataset of 264 documents, classified by domain experts on the basis of the TGR, to be used for validation purposes. In the Gold Dataset, the documents are distributed as follows:

Table 18 – Distribution of documents in the Gold Dataset within the TGR

Class	Number of documents
1 - Governing Bodies, Legal-Institutional Affairs, and Communication	10
2 - Organization, Assets, and Instrumental Resources	19
3 - Human Resources	21
4 - Financial Resources, Accounting, and Tax Management	15
5 - Information Systems	9
6 - Programming, Coordination, and Control	41
7 - Agriculture and Livestock	35
8 – Handicrafts	3
9 – Trade	5
10 - Industry and Extractive Activities	6
11 - Tourism and Accommodation Facilities	4
12 - Territorial Planning, Urban Planning, and Construction	6
13 - Infrastructure and Transport	8
14 - Environmental Protection	15
15 - Health, Hygiene, and Veterinary Medicine	28
16 - Social Policies	7
17 - Education, Training, and Labor	11
18 - Cultural Heritage and Activities	15
19 - Sports and Recreational Activities	6
TOTAL	**264**

From Table 18, it can be observed that the classification entries present in the dataset are not uniformly distributed, which is why the idea of using them to train a model was discarded. Further analysis showed that data belonging to certain classes, such as class 8, are scarcely represented. This may be due to the wide variety of topics covered by the TRER classes, some of which may be popular in certain Italian regions and rare in others.

3. 2.2. Text preprocessing

The processing of the data presented several challenges. The main issue stems from the length of the documents, which average around 3,000 words, the lack of structure, which varies according to the author of the document, and the fact that they are written in bureaucratic language, making it impossible to automatically cut out sections of text that are irrelevant for classification. The absence of a consistent and stable structure limits the applicability of segmentation techniques based on textual patterns or formal markers. Furthermore, the text contains numerous references to laws and legislative decrees without explicit mention of their content. These citations, used to justify administrative decisions, occupy significant portions of the text but do not provide useful information for the correct classification of the document, causing the relevant information to be hidden within text that appears meaningless to those who are not experts in the field.

The analysis of administrative documents in PDF format highlights several critical issues regarding their automatic classification. Although the first page contains structured data such as the proposer and the subject, these are often unreliable: the proposer can mislead the system (e.g. acts from the Department of Digitalization labelled as "Infrastructure"), while the subject, lacking standardization and often limited to internal identification codes, risks being misleading or of little significance for the intended purposes.

With regard to the main body of the text, the decision section (following the word "Determina") initially appears to be the "cleanest" as it is free from legislative citations. However, focusing on this part risks reducing the classification to purely economic topics, since it mostly contains indications of monetary amounts or technical references to previous acts. In contrast, the premises (the portion of text preceding the "Determina") provide the most relevant thematic details. In conclusion, the best strategy is to exclude the final decision to avoid distortions and focus the analysis on the premises.

3.2.3. Experimentation

A. Unsupervised Clustering Experiments

The initial attempts to organize the documents were based on unsupervised learning techniques, aiming to group the texts into the 19 TGR categories without the aid of pre-existing labels.

- K-means and TF-IDF: the texts were converted into numerical vectors using TF-IDF, a statistical method used in natural language processing and information retrieval to evaluate how important a word is to a document in relation to a larger collection of documents[17]. The k-means algorithm was then applied. It is one of the most popular "unsupervised" machine learning algorithms used for clustering. Its goal is simple: group similar data points together and discover underlying patterns without the need for pre-existing labels[18]. The results were unsatisfactory: the complexity of the data and the subtle semantic distinctions between the numerous classes meant that the centroids were unable to accurately separate the documents;

- Lbl2vec: it leverages semantics to guide the clustering process toward categories that are meaningful to the user[19]. Despite the introduction of manual descriptors, the experiment failed: only 41 out of 264 documents were classified correctly. The intrinsic similarity between administrative categories confirmed that methods based solely on word vectors are not a viable solution.

B. Zero-Shot Classification Experiments

Attention has shifted toward zero-shot classification, a paradigm where models – specifically Large Language Models (LLMs) – must assign categories to text based solely on linguistic descriptions. This allows them to classify inputs into categories they were never explicitly exposed to during the training phase. Several language models were tested, yielding divergent results:

- Small Models (BART, Distil-RoBERTa, T5): they demonstrated extremely poor performance, even worse than Lbl2vec, highlighting that for such specific tasks, the scale of the model and the depth of training are crucial;
- Large Language Models (LLM): leading models such as GPT-4o, GPT-4o-mini, Llama 3-8B, and Gemini 1.5 Flash were compared.

[17] Understanding TF-IDF (Term Frequency-Inverse Document Frequency), <https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/>

[18] K-Means Explained, <https://towardsdatascience.com/k-means-explained-10349949bd10/>

[19] Unsupervised Text Classification with Lbl2Vec, <https://towardsdatascience.com/unsupervised-text-classification-with-lbl2vec-6c5e040354de/

Comparative results:

- GPT-4o: This model proved to be markedly superior. It demonstrated a clear ability to handle long prompts (essential for the analysis of entire administrative acts) and to accurately interpret the nuances of the 19 classes;
- Llama 3-8B: It turned out to be the least effective. In addition to slow operation, it showed difficulties in following the required output formats and, above all, a poor ability to manage long texts, making it incompatible with the nature of the PDF files under consideration;
- Gemini and mini versions: These occupied a middle ground, confirming that large-scale generalist models offer the best guarantees for the classification of complex administrative documents.

The results of zero-shot classification suggest that the complexity of administrative acts requires a level of semantic understanding that only large-scale Large Language Models possess. As shown in the table below, it is clear that lighter models and traditional clustering methods fail because they cannot distinguish the subtle nuances between the 19 TGR categories, which are often characterized by similarly bureaucratic language. A decisive factor for success is the management of the context window: models that do not support long prompts, such as Llama, prove inadequate since the documents require the analysis of extensive preambles to extract their true subject. By contrast, the excellence of GPT-4o shows that the scale of the model and the depth of training enable it to overcome the limitations of statistical text analysis, making zero-shot classification the only practical solution. Ultimately, success depends on the model's ability to accurately interpret the natural language descriptions of the classes, isolating the conceptual content from distortions caused by monetary figures or regulatory references.

Table 19 – Results of the models in Zero-shot classification

Model	Macro F1-Score	Accuracy
BART	0.04	0.08
T5	0.09	0.11
RoBERTa	0.06	0.06
GPT 4o	0.67	0.65
Gemini	0.44	0.43
Llama	0.15	0.19

The F1-score (or F-measure) is a statistical metric used to evaluate the accuracy of a classification model. It is particularly useful when dealing with imbalanced datasets, where some classes are significantly more frequent than others. The Macro F1-score treats all classes equally, regardless of the number of documents they contain.

C. Prompting Experiments

The classification of managerial acts using AI requires advanced prompting strategies to overcome the ambiguity of bureaucratic language. Four main methodologies have been tested:

1. One Call: a single prompt including all 19 classes and a guiding example. It is fast and economical, but it risks overloading the model;
2. Multiple Calls: the task is divided into 19 calls (one per class). The model assesses the document's belonging to each individual category. It is accurate but extremely costly and slow;
3. Two Steps: a two-stage process in which the most probable classes are first selected and then, among these, the final one is chosen. Excellent for resolving semantic overlaps;
4. Second Chance: after an initial choice, the model is asked to review its decision to correct any possible errors.

Testing all the techniques revealed that the first and third were the most effective. The first was chosen because it requires only one call per document, making it the fastest and most cost-effective solution.

D. Creation of a Silver Dataset

After identifying the optimal prompting technique, the research continued with the aim of creating a training set for a model dedicated to the classification of administrative acts that would be quick and executable anywhere, allowing archivists to classify documents without relying on LLMs.

1.404 new managerial acts, which had not previously been catalogued and were separated from the original Gold Dataset, were automatically classified. This process led to the creation of a Silver Dataset—a fundamental resource that drastically reduces the time and cost of manual cataloguing. Having such a large volume of pre-classified data enables the training of proprietary models or the fine-tuning of existing models, without the burden of constructing entire training sets from scratch.

However, the analysis of the document distribution confirmed a significant imbalance among the TGR classes. Despite targeted research efforts, some categories remain underrepresented or entirely absent, such as Class 8. In contrast, other classes (such as Class 7) are extremely frequent, reflecting a skewed distribution already observed in the Gold Dataset and typical of the real administrative domain.

Table 20 – Distribution of documents in the Silver dataset by TGR class

Class	Number of documents
1 - Governing Bodies, Legal-Institutional Affairs, and Communication	47
2 - Organization, Assets, and Instrumental Resources	110
3 - Human Resources	26
4 - Financial Resources, Accounting, and Tax Management	45
5 - Information Systems	28
6 - Programming, Coordination, and Control	86
7 - Agriculture and Livestock	422
8 – Handicrafts	4
9 – Trade	19
10 - Industry and Extractive Activities	29
11 - Tourism and Accommodation Facilities	21
12 - Territorial Planning, Urban Planning, and Construction	37
13 - Infrastructure and Transport	32
14 - Environmental Protection	82
15 - Health, Hygiene, and Veterinary Medicine	45
16 - Social Policies	50
17 - Education, Training, and Labor	218
18 - Cultural Heritage and Activities	53
19 - Sports and Recreational Activities	51
TOTAL	**1.404**

3.2.4. Analysis of the results

The main objective in creating the Silver Dataset is the development of a highly specialized proprietary model. This approach provides significant strategic benefits: an “in-house” model would be more compact, faster, and more efficient, drastically reducing computational costs and response times compared to generalist solutions. Moreover, technological independence would allow for the elimination of reliance on third-party services and APIs. To validate the effectiveness of the Silver Dataset as a training base, classic machine learning algorithms such as Logistic Regression (LR)[20] and Linear Support Vector Classifier (SVC)[21] were tested. At this stage, the models were trained on automatically generated data and then evaluated on the Gold Dataset (certified by experts), using the TF-IDF vector representation of metadata as input for comparison.

Table 21 – Results of the evaluation carried out by models trained on the Silver dataset on documents from the Gold dataset

Model	Macro F1-Score	Accuracy
LR	0.38	0.44
SVC	0.55	0.56

As can be seen from Table 5, in line with expectations, the proprietary models performed worse than GPT-4o used in zero-shot mode. However, comparison with other Large Language Models revealed surprising results: the SVC model not only outperformed Logistic Regression but also achieved better results than Gemini.

Even more noteworthy is the fact that both lightweight classifiers (LR and SVC) have outperformed Llama. This outcome is highly significant from a strategic standpoint, as it demonstrates that relatively simple algorithms, when trained on a well-constructed Silver Dataset, can provide a more effective and high-performing alternative to many free or open-source LLMs for specific classification tasks.

In addition to accuracy, time is a crucial factor in practical applications. The use of LLMs requires prohibitive processing times: running Llama on standard hardware took an entire day, while Gemini, even when accessed via API, required several hours. In contrast, training and testing models such as SVC or Logistic Regression can be completed in just a few minutes on the same hardware. This efficiency is vital for institutions that require frequent updates: once the Silver Dataset has been consolidated, creating a new classifier becomes an almost immediate process.

In conclusion, although GPT-4o remains the benchmark for absolute performance, the SVC model demonstrates the validity of the entire pipeline. It enables the development of lighter, more cost-effective, and competitive tools, even surpassing the results of free LLM models. This approach transforms automatic classification from an academic challenge into a practical and sustainable solution for everyday use, especially in contexts where computational and financial resources are limited.

4. Internal experimentation at PaRER

In 2024, ParER initiated a collaboration with FarNetwork srl to experiment with the possible integration of agents developed using Microsoft Copilot for document management. Although the initial tests on semantic search and summarization (using the GPT-4o model on Copilot Studio) delivered promising results, a technological limitation became apparent in July 2025: the lack of support for Service-to-Service

[20] Logistic Regression in Machine Learning, <https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/>

[21] Support Vector Machine (SVM) Algorithm, <https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/>

(S2S) interactions via API on the Copilot Studio client. As a result, and considering the findings from experiments conducted by the University of Bologna, ParER shifted its research focus towards the use of AI models for zero-shot classification of protocol records (incoming and outgoing correspondence) within its own TRER classification scheme, particularly regarding the first-level entry.

4.1. Description of Data and Domain

The dataset consists of electronic administrative documents, either incoming or outgoing, which have been acquired through protocol registration in the digital document management system of the Regione Emilia-Romagna. All documents are in Italian, predominantly born-digital, and comprise multiple files of various formats, with an average of three files per registration.

These are public records that are not available online; they have been exported from the authority's document management system to be stored exclusively within a protected folder on the corporate SharePoint, thus ensuring the highest standards of security and integrity during the experimental phases.

Models available as Azure OpenAI services were used; unlike the public version, in this case the data remains within the company's dedicated Azure tenant and is not used to train global models. This activity was formalized by updating the Record of Processing Activities (RoPA) required by Article 30 of the GDPR, in accordance with the guidelines provided for appointees and the data controller, including a specific note regarding the Copilot experimentation. Employees of the external provider were involved in the process as they operate under the authority of the Regione, which is the data controller pursuant to Article 29, paragraph 1, of the GDPR.

The documents generally consist of a main document and several attachments, with both form and content being highly variable—particularly in the case of incoming documentation, which may originate from citizens, private individuals, or other public institutions. The aim of the experiment is to classify documents received from external sources in an unsupervised manner, according to the classes of the TRER, using the zero-shot classification method. As these documents have already been registered and archived, it is possible to directly assess the results against the classifications already assigned by the expert staff of RER.

The TRER contains 21 first-level classes, each identified solely by a name, without any brief description that could serve as external knowledge for the classifier. Additionally, there are a further 11 classification entries marked by codes starting with 800, intended for the management of specific partitions for particular subjects. The hierarchy of the TRER generally extends to the third level, with some residual cases reaching as far as the fifth level. Consequently, the structure of the TRER is highly complex and consists of 1,029 distinct entries.

Table 22 shows the distribution of the selected documents within the first class of the Regione Emilia-Romagna's classification scheme. Once again, it is evident that the distribution of the selected documents is not uniform across the classes.

Table 22 – Distribution of documents in the Classification Scheme of the Regione Emilia-Romagna

Class	Number of Documents
100 - Governing Bodies, Institutional Affairs, and International Relations	266
150 - Legislative and Legal Affairs	26
200 - Organization and Human Resources	213
250 - Financial Resources	250
270 - Equipment and Assets	74
300 - Information Systems and Communication	78
350 - Planning, Coordination, and Monitoring	22
400 – Agriculture	505
430 - Development Policies and Promotion of Productive Activities	76
440 - Fisheries Economy	21
450 – Trade	17

Class	Number of Documents
460 – Tourism	53
470 - Energy Policies	20
500 - Territorial Planning	83
520 - Mobility and Transport Systems	76
550 - Environment, Soil, and Coastal Protection	288
600 - Healthcare and Social Policies	299
700 - Education and Vocational Training	28
710 - Labor Policies and Employment Services	2
720 – Culture	37
730 - Sports and Recreational Activities	7
800 - Special Entities	35
Total	**1,519**

The annual average of registered protocol entries is 1.200.000. Of these, 70% pertain to documents originating from external sources, equivalent to an annual average of 828.000 documents. Furthermore, around 70% of inbound flows are made up of automated flows (web forms or similar); the remaining 30%, amounting to approximately 250.000 documents, are manually classified by RER operators. Among these, a dataset of 1.519 already classified documents was selected.

Since it is not possible to identify a primary category for each record when multiple classifications are present, it has been necessary to account for every document under each assigned classification. This resulted in a total of 2.350 classifications to be used for validation purposes.

4.2. Text preprocessing

The selected documents are produced by external parties and transmitted to RER, where expert staff have registered them in the document management system. Consequently, as these are not documents produced directly by the RER, a content analysis comparable to the one conducted during the experimentation on managerial acts is not possible. The language used varies depending on the sender; it may be highly technical in the case of other public administrations or businesses, or less formal in the case of citizens.

The documents in the dataset were originally composed of several files in different formats, typically consisting of a primary document and its attachments, each accompanied by metadata clarifying their hierarchical relationships. During the pilot project, an attempt was made to classify sets of documents composed of multiple files grouped in a folder[22], but this yielded poor results. Consequently, a normalization process was carried out to produce a single PDF file containing all the content, organized according to these relationships. It must be noted, however, that incoming communications are generally transmitted via email. The relevant content may be located within the body of the email or in one of the attachments, and this distinction may not always clearly emerge from the document's metadata.

The documents were selected so as to exclude non-text file formats (such as image or video files). The text varies greatly in length and may include charts and images.

[22] A subsequent test conducted with GPT-5.1 was successful: it was possible to correctly classify an entire folder of files.

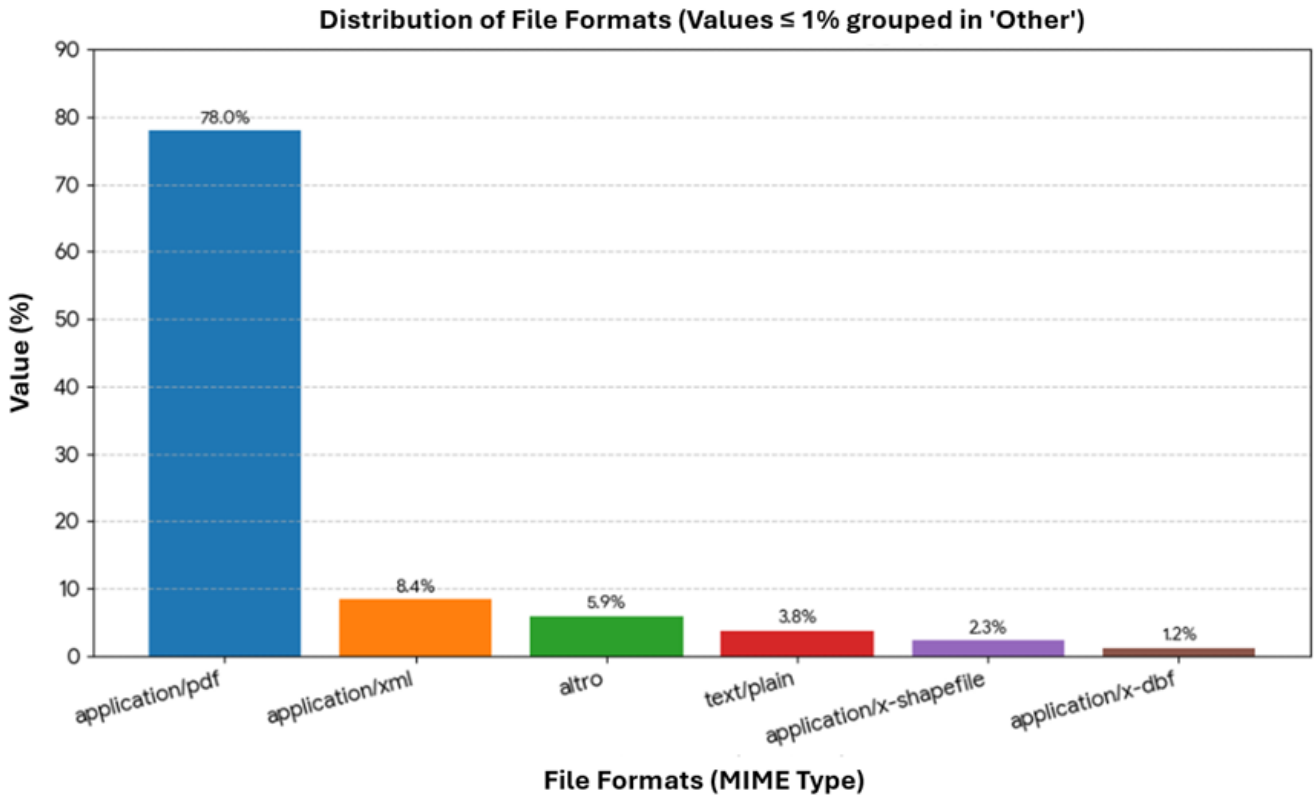

Figure 12 – Example of average distribution of file formats (by MIME type) present in the registered documentation

4.3. Experimentation

During the experimentation, a Proof of Concept (POC) application was developed to manage a complex process: downloading a PDF from SharePoint, extracting and optionally summarizing the text, classifying it using AI, and returning the results as structured output (usable by any potential calling service) to be made available to RER staff as a suggestion for classification.

The classification results are returned in JSON format, including the file name, the code of the assigned classification, its description, the AI's confidence level in assigning it, and the reasoning behind the proposed outcome. Specifically, the classification operation can be invoked by inputting the entire TRER or by using the "two-round" option, which first uses only the top level and then applies the TRER limited to all child classifications of those found in the first round.

A. Classification with Mistral SMALL 3.1

Mistral Small 3.1 can process and understand visual inputs and lengthy documents. It is a versatile model designed for various tasks such as programming, mathematical reasoning, document comprehension, and dialogue. It has been developed with low-latency applications in mind and offers best-in-class efficiency compared to models of similar quality. This model has undergone a comprehensive post-training process to align it with human preferences and needs; therefore, it is immediately ready for use in applications requiring chat or precise instructions.

Since Mistral cannot process texts exceeding 128kb (total characters: approximately 131.072, considering 128 x 1.024 bytes), a four-step procedure was developed:

1. extraction of text from PDF files using an open source library;

2. summarization of the text with an LLM model (in this case GPT-4.1, accessed via Semantic Kernel[23]) to generate a summary of the provided text;
3. classification of the summary at the first level of the file plan;
4. classification of the summary at the lower levels resulting from the first level.

The process returned 4.096 responses, with an average of approximately 2,5 responses per document. 600 documents were not classified.

B. Classification with GPT-4.1

GPT-4.1 is specifically designed to enhance coding and the execution of instructions, making it more efficient at handling complex technical and coding problems, and it can process texts up to 1 million tokens. In this case, the process is therefore simpler:

1. extraction of text from PDF files using an open source library;
2. classification of the extracted text across the entire TRER.

The process returned 1.460 responses, with an average of one response per organizational unit. 59 documents were not classified.

4.4. Analysis of the results

As shown in Table 23, the results from GPT are markedly superior to those from Mistral Small, which has a much higher number of unanswered cases (600 compared with 59) and is generally less accurate. The results of classification at the first level are fairly similar to those obtained with GPT-4o in the zero-shot experimentation on administrative acts (where the TGR has a description that can be used as a knowledge base). Nevertheless, the results are not considered sufficient for unsupervised automatic classification. The match at the third classification level, which is regarded as the reference level, is only 31%, meaning less than one correct response out of three.

Table 23 – Comparison of results between the Mistral Small and GPT models

Metrics	Mistral Small 3.1	GPT 4.1
Coverage (% of classified documents)	60% (600 not categorized)	96% (59 not categorized)
Precision on Classified Docs (Exact Match)	16%	24%
Precision on Classified Docs (Level I Match)	46%	74%
Precision on Classified Docs (Level II Match)	35%	52%
Precision on Classified Docs (Level III Match)	19%	31%
Average Number of Classes Returned	media 3,3 sigma 7,1	1
Qualitative Notes	More cautious but often "silent"; higher recall at Level II.	Responds almost every time; higher overall accuracy in exact matches.

[23] https://learn.microsoft.com/en-us/semantic-kernel/overview/

However, a sample check on the incorrect responses given by the GPT4.1 model at the first level of the TRER highlighted that, in some cases, the classification assigned by the model was actually more accurate than that given by the regional operator. This demonstrates that human decisions too can be improved. For this reason, it should be considered that, in a test such as this, even the sample itself is not entirely reliable for the purpose of automatically evaluating the results.

From a quantitative perspective, the time taken by the models to analyze 1.519 UDs was approximately 7-8 hours in total. The cost of using Mistral Small was €0.45, while the cost of using GPT 4-1 was €196, averaging €0.13 per document.

Despite these considerations, given the rapid evolution of the models and the modest cost of applying them to an average annual sample of 250,000 documents, RER has decided to continue the project in 2026 with the following activities:

- studies of fine tuning and training techniques;
- experiments with other open-source models to assess the possibility of using an "in-house" model;
- integration of the classifier prototype into the document management system.

Regarding the last point, the aim is to gradually initiate, under the supervision of specialized personnel, the introduction of the automatic classification service within the document management system, targeting a selected group of key users. Authorized operators will be able to access the service and verify the responses provided, giving a measurable assessment of their quality.

5. Conclusions

Based on the experiments conducted by the University of Bologna and ParER, the conclusions highlight how artificial intelligence represents a promising, yet not fully autonomous, resource for the classification of public documents. Table 24 presents a structured comparison of the two lines of research.

Table 24 – Comparison of the two experiments

Feature	**Academic experimentation with the Università di Bologna**	**Internal experimentation at ParER**
Dataset	1,728 public managerial acts (resolutions/determinations).	1,519 registry records.
Objective	Classification based on TGR - 19 Level I classes.	Classification based on TRER - 21 Level I classes.
Top Model	GPT-4o (Precision: 65%).	GPT-4.1 (Level I Precision: 74%).
Strength	Effective handling of long, bureaucratic texts due to model scale.	High coverage (96%) and management of contexts up to 1 million tokens.
Critical Issues	Textual heterogeneity and irrelevant legislative citations.	Lack of textual descriptions for TRER classification categories; heterogeneity of content.

The main difference between GPT-4o (released in May 2024) and GPT-4.1 (released in April 2025) lies in their technical specialization and memory capacity. GPT-4o is an "omni" model designed for seamless multimodality (voice, video, and text in real time), featuring a context window of 128,000 tokens. It is considered excellent for standard conversations but may "forget" details in extremely long documents. GPT-4.1 introduces a massive context window of 1 million tokens, allowing it to analyze large amounts of information, entire books, or complex documents in a single session without losing track. Furthermore, the recent release of GPT-5 (August 2025) suggests that performance results will likely exceed those

observed in these experiments, particularly due to its advanced reasoning capabilities and even more robust handling of multi-file structures.

However, analysis of the two trials highlights that document management is not merely a technological matter, but above all a methodological one.

The absence of standards and best practices in document production fundamentally undermines any attempt at automation or efficient management. Administrative documents are often unclear, lacking in transparency or structure. The lack of a consistent framework prevents the use of automatic segmentation techniques. Without proper drafting, AI struggles to identify the real subject, leading to distortions and hallucinations.

The classification plan is the logical tool by which we organize memory. When it becomes too complex, it turns into a limitation. A complex classification plan often presents categories with blurred boundaries, making it difficult to correctly decide where to place a document. The presence of thousands of distinct entries (1.029 in the case of TRER) and deep hierarchical levels (up to the fifth level in TRER) exponentially increases the likelihood of error. Furthermore, in complex contexts, many classes remain unused or are rarely populated, while others are overloaded. This imbalance makes both machine learning and correct manual selection by operators difficult, as the outcome of the checks has demonstrated.

The absence of supporting tools, such as ontologies providing an explicit knowledge base, forces both the system and humans to rely on subjective interpretations, which in turn generate inconsistency within the archive. AI should therefore be regarded as critical aid for expert staff, automating classification suggestions while retaining the necessity for human oversight to ensure the integrity of the archival system.

The best practices that apply to humans also apply to AI. A disorderly archive and an ambiguous classification plan produce poor results regardless of the power of the software used. AI can support logical organization, but the quality of the archival system and its understanding depend on discipline in document creation and the clarity of management tools.

12.4 Centro Italiano Studi Ufologici (CISU)

In February 2023 Team CU05 asked CISU (Centro Italiano Studi Ufologici – “Italian Center for UFO Studies – headquartered in Turin, Italy) to take part in a case-study intended to test the capabilities of AI-based technology to manage archival aggregations and enrich the metadata elements of records. To be noted that one of the members of Team CU05 – Massimiliano Grandi – is also a member of CISU.

In February 2023 the Executive Committee of CISU approved the proposal of participating in the case study in the capacity of the partner providing the documents to be processed through AI.

CISU was established in 1985. As to the size of its holdings, CISU possesses the second-largest archives in Europe (after the one kept by the Swedish association “Archives for the Unexplained” – AFU) and one of the ten largest archives in the world. The archives is structured in various series and includes personal archives of former UFO researchers: you can find in it records relating to ca. 43,000 UFO sightings reported in Italy from 1900 to date. The association does not take a particular stand as far as the explanation of the sightings is concerned but commits to following a scientific methodology or – when that is not possible for lack of data – at least a zetetic approach. The documents making up the archives are mostly in paper format, but a program to digitize them has been started. CISU has recently set up a server where both digital reproductions of paper documents and born-digital documents are kept.

As it was not possible to make available the series of the reports produced by CISU members following the investigation on specific UFO sightings – as they contained too much personal and sensitive data – it was decided to select for the case-study the large series of ca. 100,000 news clippings of publications from all over the world (but mainly from Italy), whose dates range from the 40s of the 20th century to date.

The association is entirely run on a volunteer basis and therefore there are no detailed finding aids describing the archival fonds and series: in this respect, the case-study might have been a good testing ground to see how AI can help to gather and organize metadata elements in an environment where they are non-existing or – at best – only very basic.

The first company selected to provide the AI-based application for the case-study was – in March 2023 – expert.ai, an Italian company (headquartered in Modena) that had taken part in the survey questionnaire in 2022. expert.ai has developed an application named “expert.ai Discovery”, using unsupervised learning methods and based on a technology aiming at providing a deep semantic understanding of text. However, expert.ai was only willing to process 3,000 records (out of the 100,000 making up the news clippings series of CISU) and asked Team CU05 and CISU to pay a 15,000 Euros fee to run expert.ai Discovery. As neither InterPARES Trust AI nor CISU were willing to give expert.ai such a sum of money, this attempt of cooperation fell through.

The second company requested (in May 2023) to take part in the case-study as the AI technology supplier was “Grupo Adapting” (that – like expert.ai – had taken part in the questionnaire survey in 2022 and whose main offices are located in Valencia, Spain and Barranquilla, Colombia), but – after weighing up the proposal for a couple of months – finally they turned down our invitation in July 2023.

The third business invited in September 2023 by Team CU05 to support the case-study as the technology partner was “Anzyz Technologies AS”, a Norwegian company based in Grimstad. In the same way as expert.ai and Grupo Adapting had done, also Anzyz participated in the 2022 questionnaire survey. Anzyz has developed Corpus Cube Linguistics (CCL™), an application based on Natural Language Processing and combining supervised, unsupervised and rule-based learning with a view to achieving a very high-level accuracy with significantly less data input than that needed by more standard machine learning technology.

Anzyz accepted to support the case-study for free, and the first preliminary meetings attended by Anzyz, CISU and Team CU05 were held in Autumn 2023. CISU and Team CU05 agreed in September 2023 on a list of requests to be submitted to Anzyz and including:

1. Indexation of contents (by extracting a given set of information elements such as name of people, name of places and dates).
2. Categorization of the news clippings by a pre-defined set of types (e.g. news concerning UFO sightings; general comments; book and film reviews).
3. Identification of the name and date of a publication, from indications reported on the scanned copy and consequent renaming of the digital file of the scanned copy itself.
4. Identification and extraction of the title of the publication containing the news clipping.
5. Identification of duplicates of the same item.

In October 2023 Anzyz replied that Corpus Cube Linguistics could meet requests no. 1 and 5 and – if CISU could create "concepts that return classes of documents", which was done by categorizing – as an example for Anzyz's AI-based application – more than 3,000 news clipping, also request no. 2.

CISU initially provided Anzyz with a first sample of 89 news clippings made up of PDF files which were processed through OCR by Anzyz and fed to Corpus Cube Linguistics. The first outcomes of this initial sample were shown to CISU and Team CU05 and it was noted that Corpus Cube Linguistics could work even if the OCR had not succeeded in reading all the characters (which is a likely case, as the paper news clippings are often in poor conditions). However, CISU and Team CU05 were told that Corpus Cube Linguistics – in order to be appropriately trained – would need a testbed ideally made up of 100,000 items. The more items you can feed to Corpus Cube Linguistics, the better results you will get. That would have been a problem as 100,000 items (i.e. 100,000 news clippings) was the size of the whole series made available by CISU, and therefore to enable Anzyz's application properly we would have needed a testbed as large as the series to be analysed. Anzyz also observed that 30,000-40,000 items for the training might have been sufficient for the case-study, although in this case Corpus Cube Linguistics would have performed less brilliantly.

However, another circumstance emerged and led to the definitive dismissal of the case-study, as in November 2023 Anzyz said that they were willing to work for free as professionals, but did not have an appropriate server available for the case-study, which required a huge amount of computing power: Anzyz did not possess such a kind of server and, moreover, the ones they had were taken up by their customers. They asked CISU and Team CU05 to provide either a server with the specified features or the money to rent one. As CISU was unable to meet any of the requests from Anzyz, Team CU05 asked the Executive Committee of InterPARES Trust AI to intervene and support the case study. The Executive Committee replied that it was not possible to grant Anzyz none of the above-mentioned requests, although Dr. Hrvoje Stancic – another researcher of InterPARES Trust AI – volunteered to ask the University of Zagreb he belongs to whether a spare server could be set apart for the case-study, provided Anzyz could join InterPARES Trust AI as a partner, which Anzyz did in February 2024.

A series of meetings involving Anzyz, Team CU05 and the University of Zagreb were held between March 2024 and July 2024, but the process of enabling Anzyz to use one of the servers of the University proved to be an extremely cumbersome administrative and technological work, up to the moment when Anzyz lost any interest in its involvement in the case study, and therefore in August 2024 the case-study was definitively shelved without achieving none of the goals that had been set by CISU and Team CU05.