Addressing the problem of file formats obsolescence using AI: the methodology proposed by Emilia-Romagna Digital Archival Centre

Stefano Allegrezza[1,*,†], Cristiano Casagni[2,†] and Marianna Tascone[2,†]

[1]University of Macerata, Via Crescimbeni, 30/32 - 62100 Macerata, Italy
[2]Emilia-Romagna Digital Archive Centre (ParER), Viale Aldo Moro, 44 - 40127 Bologna, Italy

Abstract

The rapid proliferation of digital technologies over the past four decades has fundamentally transformed the way information is created, disseminated, and preserved. Digital documents, audiovisual materials, datasets, and born-digital records now constitute an essential part of the cultural, administrative, and scientific heritage of contemporary societies. Yet, this unprecedented reliance on digital formats has introduced a set of preservation challenges that differ substantially from those associated with analogue media. Among these challenges, the obsolescence of file formats has emerged as one of the most pressing concerns for archivists, librarians, memory institutions, and digital preservation practitioners. The risk of format obsolescence affects all kinds of digital repositories, including digital libraries and archives. As software ecosystems evolve and proprietary standards proliferate, previously common formats may become unreadable, unsupported, or dependent on environments that no longer exist. This dynamic, if left unaddressed, threatens the long-term accessibility, authenticity, and usability of digital information.

According to Annex 2 of the "Guidelines for the Creation, Management and preservation of electronic records" issued by the Italian Agency for Digital Italy (AgID), file formats must be assessed by calculating an interoperability index, a metric that gauges their level of obsolescence and helps determine whether migration is required. In 2025, the Emilia-Romagna Digital Archive Centre (Polo Archivistico della Regione Emilia-Romagna – ParER) experimented a technical solution based on artificial intelligence to automatically calculate the interoperability index of file formats. The work was carried out with the objective of building a prototype catalog of file formats that combines authoritative knowledge from the Library of Congress (LoC) with machine-learning-based evaluations. This paper aims to present the promising results achieved.

Keywords

Artificial intelligence, conversion, digital repositories, file formats, interoperability, LoC, machine learning, migration, obsolescence, sustainability factors, text classification

1. Introduction

File format obsolescence arises from the inherent instability of the digital preservation environment. Unlike physical media, whose deterioration is largely predictable and material in nature, digital objects exist within complex technological infrastructures that include software, hardware, operating systems, codecs, and interpretive frameworks [1]. Any change or failure within this interconnected chain may render a file partially or completely inaccessible. The problem is exacerbated by the commercial logic underlying many digital tools: vendors may discontinue support for older versions of their products, modify file specifications, or close once-open standards. Even open formats can evolve, creating versioning issues that challenge long-term stewardship. As a result, digital preservation must contend not only with bit-level stability but also with the continued intelligibility of the formats in which information is encoded.

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy
*Corresponding author.
†These authors contributed equally.
stefano.allegrezza@unimc.it (S. Allegrezza); cristiano.casagni@regione.emilia-romagna.it (C. Casagni); marianna.tascone@regione.emilia-romagna.it (M. Tascone)
0000-0002-7319-2483 (S. Allegrezza); 0009-0009-7572-3725 (C. Casagni); 0009-0002-1535-6379 (M. Tascone)

In response to these risks, a wide range of strategies has been developed in the field of digital preservation. Among them, file format migration — or format conversion — has become one of the most widely adopted and debated approaches [2]. Format migration refers to the process of transforming digital objects from an at-risk or obsolete format into a more stable, open, or widely supported one. This strategy seeks to maintain ongoing accessibility by aligning objects with prevailing technological standards. Migration may be performed once (a "one-time conversion") or periodically as part of a broader digital preservation plan. While migration offers a pragmatic solution to the problem of declining format support, it also raises complex questions about authenticity, metadata integrity, file behavior, and the potential loss or alteration of essential characteristics. Consequently, the implementation of migration workflows requires meticulous planning, documentation, and quality assurance.

Despite longstanding awareness of the risks posed by format obsolescence, the pace of technological change poses new challenges. The increase in proprietary cloud-based platforms, the explosion of multimedia formats, and the rise of artificial intelligence–generated content have added further layers of complexity. Many modern digital objects are composed of intertwined components — such as embedded scripts, dynamic content, dependencies, or interactive structures — that cannot be preserved through simple format transformation. These developments have shifted the discussion from a solely technical matter to a broader reflection on digital stewardship, records lifecycle management, and institutional responsibility. What must be preserved: the content, the structure, the behavior, or the user experience? What constitutes an acceptable degree of alteration during migration? To what extent can institutions rely on emulation or virtualization as alternatives or complements to migration?

At the same time, global initiatives and international standards have contributed to the conceptual and operational framing of format management in long-term preservation. Models such as the Open Archival Information System (OAIS) [3] emphasize the role of representation information, documentation, and preservation planning in ensuring that digital objects remain understandable to designated user communities over time. Numerous guidelines and registries — such as PRONOM [4], the Library of Congress Sustainability of Digital Formats [5, 6], the National Archives of United Kingdom Guidelines [7], etc. — offer tools to assess risks, identify appropriate target formats, and evaluate the sustainability of different file types. These instruments have become essential for institutions facing ever-increasing volumes of digital materials requiring continuous oversight and strategic decision-making. Nonetheless, the existence of such tools does not eliminate the need for sustained human judgment, organizational coordination, and clear policies governing when and how to undertake format migration.

The concept of obsolescence must also be understood in its broader socio-technical dimension. Formats do not become obsolete solely due to technological decay; they become obsolete because communities abandon them, practices shift, and new forms of digital communication emerge. Format obsolescence is therefore a socio-technical phenomenon shaped by market dynamics, institutional choices, user behavior, and cultural expectations. This means that effective preservation strategies cannot rely solely on technical fixes: they require ongoing monitoring of technological trends, engagement with software and standards communities, and proactive data governance. In other words, format migration is not merely a reactive response to technological decay; it is an integral component of the long-term stewardship of digital heritage.

2. The interoperability assessment proposed by the Agency for Digital Italy

Among the many methodologies proposed for assessing the obsolescence of file formats [8, 9, 10, 11, 12, 13, 14, 15], particularly interesting is the one developed by the Agency for Digital Italy (in Italian: Agenzia per l'Italia digitale – AgID) in its "Guidelines on Creation, Management and Preservation of Electronic Documents" (in Italian: "Linee Guida sulla formazione, gestione e conservazione dei documenti informatici"), published on September 9, 2020 [16]. Among these innovations, the concept of "interoperability assessment" — explored in detail later — stands out as a key step toward ensuring both efficient electronic records management and sustainable long-term preservation.

A significant part of this effort is found in Annex 2, which is entirely dedicated to file formats and migration strategies. This annex proposes a structured, though partly subjective, approach for selecting suitable formats. It is divided into three main sections. The first, "Introduction", contains foundational concepts and a taxonomy of file formats. The second, "File formats", is the main part of the Annex and describes 124 file formats divided into 16 categories:

1. paginated documents;
2. hypertexts;
3. structured data;
4. email;
5. spreadsheets and multimedia presentations;
6. raster images;
7. vector images and digital modelling;
8. fonts;
9. audio and music;
10. video;
11. subtitles, captions and dialogues;
12. containers and multimedia packages;
13. compressed archives;
14. administrative documents;
15. applications and source code;
16. cryptographic applications.

The third, "File formats Recommendations", outlines strategies and recommendations for format migration. This section proposes a quantitative methodology for evaluating file formats, identifying those that are in danger of becoming obsolete and choosing those that are more likely to be preserved. To apply this method a group of nine criteria (Figure 1) — each of which is assigned a numerical value (score) [17, 18] — must be assessed:

a) **Standardization** (from 0 to 3 points). Formats recognized as *de jure* standard by a standardization body (such as ISO, UNI, W3C, CEN, ITU, SMTPE, etc.) are awarded with the maximum score, equal to 3; those that have not received this recognition but have become *de facto* standard, thanks to their widespread diffusion, achieve a score of 2. Finally, a format can also be neither *de jure* nor *de facto* standard (score equal to 0), but in this case you should tend to exclude it from the list of acceptable formats.

b) **Disclosure** (from 0 to 3 points). Open formats, *e.g.* those whose specifications have been published and made available — possibly also for a fee — obtain the maximum score, equal to 3; on the contrary, closed formats, *e.g.* those whose specifications have not been made available, achieve the minimum score, equal to 0.

c) **Proprietary** (from 0 to 4 points). Non-proprietary formats, *e.g.* those that are not encumbered by intellectual property rights and whose specifications are not managed by a private organization but by a community of developers (for example, the LibreOffice community) or by a standardization body, achieve the maximum score, equal to 4. Proprietary formats, *e.g.* those that have been created by a private organization (for example, a software house), which owns the intellectual property rights and manages the specifications, achieve a variable score, ranging from 0 to 3. In particular, formats that are proprietary but free to use achieve a score of 3. Proprietary formats who allow the reading of documents already encoded in this format but not the production of new ones achieve a score of 2. Finally, those that do not even allow the reading of documents encoded according to this format achieve a score of 0.

d) **Extensibility** (from 0 to 2 points). Extensible file formats, *e.g.* conceived from the outset to allow subsequent versions that progressively increase their functionality, achieve the maximum score, equal to 2; conversely, those that cannot be extended receive a score of 0.

Figure 1: Criteria to evaluate when calculating the interoperability index.

e) **Level of metadata** (from 0 to 3 points). It takes into account the level of connection between the document and the metadata associated with it. Formats that allow you to embed metadata within the record get the highest score of 3.

f) **Robustness** (from 0 to 2 points). Non-robust file formats, *e.g.* those that, in the case of corruption of the bit stream (the sequence of "0" and "1" of which an electronic record is made up), do not allow to recover any part of the original file, achieve the minimum score, equal to 0. Completely robust file formats, which include mechanisms to verify any loss of integrity and allow the recovery of the entire document (or, in the case of container formats, of the intact parts), achieve the maximum score, equal to 2. Partially robust file formats achieve a score equal to 1.

g) **Device independence** (from 0 to 4 points). File formats that are independent from the device, *e.g.* can be represented in a reliable manner and always in the same way independently from the hardware platform and the software used, achieve the maximum score, equal to 4; those dependent on the device achieve the minimum score, equal to 0.

There are also two other evaluable criteria (to which, however, Annex 2 does not assign a specific score):

h) **Backward and forward compatibility**. Backward compatible file formats are those you can use with one of the previous versions of the software that produced them; forward compatible formats are those you can use with any software version subsequent to the one that produced them.

i) **Textual or binary encoding**. Textual formats are those that allow you to extract the information content by encoding each byte of the binary sequence with the corresponding character (for example, in the ASCII code); conversely, formats that do not allow this type of representation are binary.

Based on this assessment, each of the nine evaluable criteria can be assigned a numerical score (except for the last two criteria, which Annex 2 does not associate with specific values). The Agency for Digital Italy refers to the sum of these scores as the "interoperability index," although many consider this terminology somewhat misleading. The index ranges from 0 to 21: a format that reaches a value of 21 is considered fully interoperable, while a score of 0 indicates the lowest level of interoperability. A threshold of 12 is regarded as the minimum acceptable level; formats scoring 12 or above are considered "acceptable," whereas lower values signal concrete issues that should be promptly addressed, for instance through migration strategies or other remedial approaches.

Special care is required when assessing container file formats. In such cases, the evaluation must consider not only the container itself but also every digital object stored within it, applying the lowest — meaning the least favorable — score among them. For example, a multimedia container requires separate assessments of the audio and video formats it contains. Similarly, a package file format must be evaluated by considering all the digital objects included in the package; if the package nests other container formats, these too must be assessed according to the same principle.

Although this methodology strives to be as objective as possible, a certain degree of subjectivity remains: if the interoperability index is calculated by different parties, it is possible that different values will be obtained. To resolve this issue, the calculation should be based on parameters that are as objective as possible, for example by drawing information from reliable registers and using techniques that leave little room for subjectivity. This will be discussed in the next chapter.

3. The methodology proposed by the Emilia-Romagna Digital Archival Centre

3.1. Project Background

The Emilia-Romagna Digital Archival Centre (in Italian: Polo archivistico della regione Emilia-Romagna — ParER) [19] is a trusted digital repository of Emilia-Romagna region (Italy), responsible for the permanent 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. Recognising the growing complexity of assessing the long-term sustainability of digital formats, in 2023 ParER initiated a collaborative project with the AI & Data division of Engineering Group [20]. The aim of this initiative was to employ artificial intelligence to compute the interoperability index. The project was completed in 2025 with the production of two deliverables: the "Study for the evaluation of a format interoperability" (in Italian: *Studio per la valutazione di un indice di interoperabilità dei formati*) and the "Technical Solution for Calculating the Interoperability Index of Digital Formats" (in Italian: *Soluzione tecnica per il calcolo dell'indice di interoperabilità dei formati digitali*). These works aimed to redefine how institutions could evaluate the interoperability of digital formats on a large scale.

The study and the technical solution were first presented on 16 May 2025 during the seminar "Archives and Responsible Artificial Intelligence. Experiences, Opportunities, Cautions" (in Italian: "Archivi e intelligenza artificiale responsabile. Esperienze Opportunità Cautele") [21], and subsequently on 10 June 2025 within the webinar "Artificial Intelligence in Archives: Solutions in Practice and Concrete Cases" (in Italian: "L'intelligenza artificiale negli archivi: soluzioni in campo e casi concreti") organized by Formez [22].

The technical solution developed by ParER represents an attempt to combine authoritative external sources with the interpretative capabilities of generative Artificial Intelligence. Although the Italian AgID framework (Annex 2) provides a structured set of interoperability criteria, the growing number of formats and their rapid evolution make manual, expert-driven evaluation increasingly difficult to sustain. The prototype explores whether a large language model (LLM) can address this challenge by analyzing textual metadata taken from external repositories and transforming it into reliable, explainable interoperability ratings.

3.2. Foundational Principles and Data Sources

Preservation planning relies on understanding the characteristics of formats embedded in digital archives, ranging from disclosure, transparency, and adoption to more nuanced aspects such as patents, hardware dependence, or self-documentation. AgID formalizes these elements into a series of criteria, each associated with specific classes and corresponding scores.

The project grounded AgID requirements on the descriptive tradition of the Library of Congress (LoC) Format Register [5], which offers rich narratives on sustainability for over five hundred formats. The decision to anchor the model in LoC data is based on the following principles:

- **International Authority**: the LoC format sheets are among the most authoritative worldwide, and their preservation-oriented vocabulary resonates strongly with the requirements of long-term digital stewardship.
- **Data Enrichment**: the LoC was selected as the primary source not only for its authority but also because it offers a rich data structure, including the possibility of finding matches with external registries such as PRONOM [23] and Wikidata [24] (starting around 2018), facilitating access to multiple sources for research and validation of technical specifications [25].

The project positioned itself at the intersection of two distinct domains: the regulatory framework defined by AgID, characterised by its rigid categorical prescriptions, and the descriptive approach of the LoC repository, which provides detailed information but refrains from explicit classification. ParER's internal registry—featuring expert interoperability assessments for 187 file formats—served as the Gold Standard against which the project's results were validated.

3.3. Materials and Methods

The entire development process followed the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining), ensuring a structured and reproducible data science framework. Creating a unified catalogue required a substantial amount of preliminary work to bridge the conceptual gap between the LoC's Sustainability Factors and the AgID's interoperability criteria. The process began with the extraction of relevant sustainability passages from 571 LoC XML files (by sustainability factors, 3976 total records), followed by their normalization and enrichment. Some aspects are worth highlighting. The primary issue was the conceptual discrepancy between LoC's sustainability factors and AgID's interoperability criteria, which require a classification into discrete scores (0, 1, or 2). Also, for these reasons a partial mapping was defined (Table 1). For the "Standardization" criterion, automatic detection of formal standards (ISO, W3C, OASIS, RFC) within the Documentation section was implemented to guide the "de jure" scoring, because the factor "Adoption" was not sufficient. Due to the lack of clear correspondence in the LoC narratives, two AgID criteria were either excluded or proved problematic: the "Extensibility" (d) criterion was excluded because not reliably mappable; the "Backward and Forward compatibility" (h) criterion was excluded due to its subjective nature and difficulty in automated analysis. Furthermore, the LoC "Transparency" factor was excluded because not reliably mappable. Furthermore, textual LoC evaluations underwent pre-processing to handle ambiguities (*e.g.*, "None identified", "unknown"), empty strings, and internal references to other records (*e.g.* `<fddLink id=>`).

3.4. The Generative AI Model Implementation

Rather than adopting a rigid rule-based mechanism or attempting to translate LoC narratives into structured metadata, the project leveraged a multilingual Large Language Model (LLM) capable of semantically understanding complex textual descriptions. The model was given the AgID scoring classes as contextual constraints (via System and Human Templates). The system effectively asked the model to "read" each LoC sustainability factor description, reason about its implications, and assign the appropriate AgID class. This approach allowed the model to handle differences in vocabulary, implicit cues, and textual nuances that would be difficult to capture algorithmically.

The decision to integrate both Italian and English sources without machine translation introduced an additional layer of robustness, leveraging the model's inherent multilingual understanding. At the end the model produced a structured output for each criterion, including:

- the AgID Class (*e.g.*, "open," "limited proprietary");
- the numerical score (0, 1, or 2);

Table 1
Partial mapping of AgID criteria onto LoC factors.

AgID criterion	Score	LoC factor	LoC factor description
Standardization (a)	0,2,3	Adoption	The extent to which the format is already used by major creators, distributors or users of information resources. This includes its use as the primary format, for distribution to end users, and as a means of interchange between systems.
Disclosure (b)	0,3	Disclosure	The extent to which comprehensive specifications and tools exist and are accessible to those who create and maintain digital content to validate technical integrity. Long-term preservation of digital content is impossible without understanding how information is represented (encoded) as bits and bytes in digital files.
Proprietary (c)	0,2,3,4	Impact of Patents	Presence of patents and therefore licensing terms.
Extensibility (d)	0,2	–	–
Level of metadata (e)	0,1,2,3	Self-documentation	The ability of a digital format to contain (in transparent form) metadata beyond that required for simple display of the content in today's technical environment.
Robustness (f)	0,4	Technical Protection Mechanism	Appropriate measures to preserve digital content and make it accessible to future generations.
Device independence (g)	0,4	External dependencies	External dependencies refer to the degree to which a particular format depends on specific hardware, operating system, or software to be displayed or used, and the expected complexity of managing these dependencies in future technical environments.
–	–	Transparency	Transparency refers to the degree to which the digital representation is open to direct analysis with basic tools, including human readability using a text editor.
Backward and forward compatibility (h)	0/?	–	–
Textual or binary encoding (i)	0/0	–	–

- a Motivation (rationale) in natural language justifying its decision. This motivation serves as a critical diagnostic tool, ensuring the traceability of the evaluation process.

The final interoperability index was computed as the arithmetic sum of the scores assigned to the six mapped criteria (maximum potential score of 20). The implemented process (Figure 2) involves calculating the AgID criterion class based on the dataset obtained from the LoC extraction (normalised through pre-processing). Using a structured prompt, the model generates an evaluation relating to the AgID class, to which the score corresponding to the class is then assigned. Finally, a post-processing operation creates a structured response that forms a row in our catalogue.

3.4.1. The prompt

The core of the classification system relied on a structured communication mechanism where the final input to the Large Language Model (LLM) was not a single command, but a dynamic combination of two distinct templates: the static System Template and the dynamic Human Template. This design ensured that every evaluation step was performed under strict normative constraints.

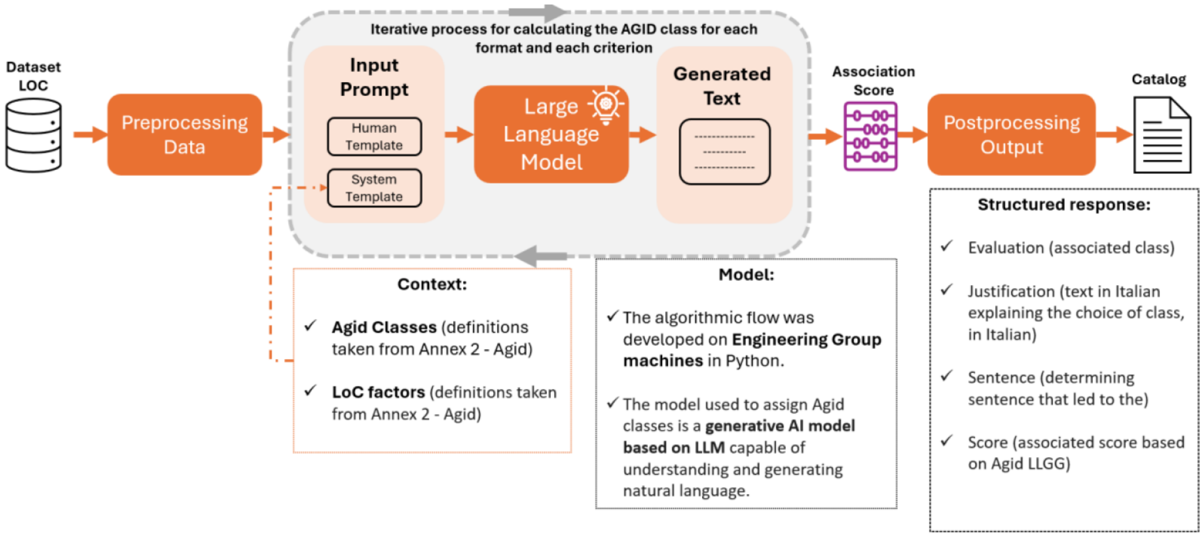

Figure 2: Workflow schema.

Table 2
Human template schema.

Human template	Parameter name	Description
Request	Textual evaluation LoC factors	Textual evaluation of the LoC factors (in English), extracted and preprocessed, to which a class from AgID Annex 2 is associated
	AgID Criterion	Name of the AgID criterion, obtained from the partial mapping, on which to perform the class assignment
Context	AgID score classes	Description of the score classes in relation to the criterion to be evaluated

3.4.2. The Human Template (The Dynamic Data Input)

The Human Template was the dynamic component responsible for feeding the unique data point for the specific classification being performed. In every iteration of the workflow, the Human Template was dynamically generated to include three primary pieces of information: first, the explicit Target Criterion (*e.g.*, "Proprietary" (c)) to focus the model's attention; second, the raw, descriptive LoC Textual Evidence (the extracted text from the Sustainability Factor of the current format); and third, a clear Action Command instructing the model to process this specific evidence against the global rules defined in the System Template (Table 2).

3.4.3. The System Template (The Fixed Rulebook)

The System Template (Table 3) served as the fixed rulebook and established the model's environment and constraints. It was the crucial component that defined the classification rules derived from the AgID Annex 2 framework [17]. First, it defined the model's persona, instructing it to act as an expert analyst in digital preservation. Second, and most importantly, it detailed the entire normative taxonomy. This included explicitly listing all the AgID criteria (*e.g.*, "Proprietary", "Disclosure", "Robustness") and their corresponding scoring rules (the exact classes like "de jure", "open", "limited proprietary") along with the associated numerical scores (0, 1, or 2). Crucially, the System Template mandated the structured output format (a JSON-like object containing the Class, Score, and natural language Motivation), ensuring the results were machine-readable for the final Interoperability Index calculation. It also provided the strict

Table 3
System template schema.

Human template	Parameter name	Description
Logical steps	Task Definition	Description of the type of problem the model is supposed to solve (in this case, the classification of the LoC text into one of the available score classes).
	Output specification	Guidance on output formatting (*e.g.*, the model should return output as a text string or more structured output, such as a dictionary). The output should contain both the evaluation and the justification
	Management of special cases	Indications to the model on how to behave with non-usable assessments (*e.g.*, Unknown to the user)

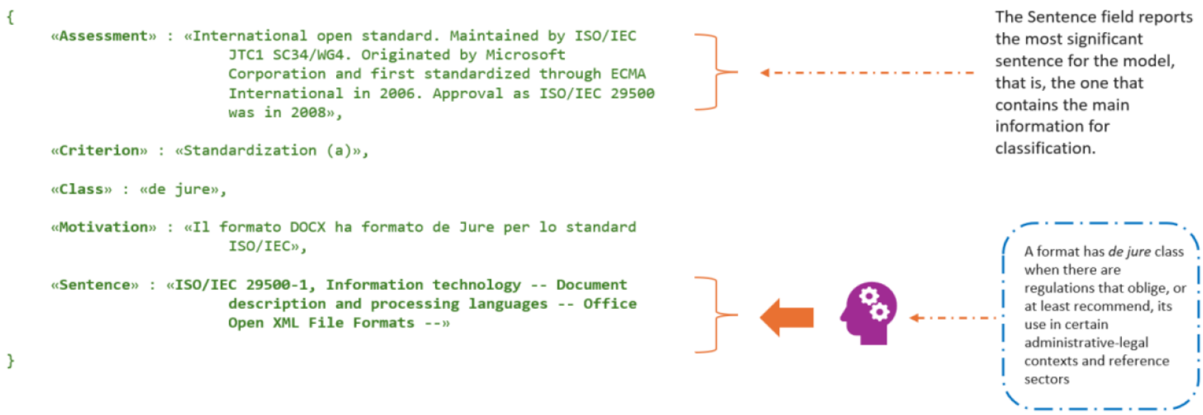

Figure 3: Example of a response to the verification of the DOCX format on the standardization criterion (a).

rule for uncertainty handling, specifying that the model must return the “I don’t know” classification if the provided textual evidence was insufficient or too ambiguous. This template remained constant for all formats and all criteria.

3.4.4. The Combined Prompt and Execution

When an evaluation was needed (*e.g.*, classifying the “Standardization” criterion for a certain file format), the two templates were concatenated. The LLM received the System Template (the rules) followed immediately by the Human Template (the data and the task). This structure compelled the model to apply its powerful semantic reasoning—its ability to understand textual nuances—to the descriptive LoC text, but only within the strict categorical boundaries and scoring logic mandated by the AgID framework, thus successfully bridging the gap between descriptive narratives and normative scores (Figure 3).

4. Results and validation

To understand how well the prototype aligns with expert judgment, a subset of 25 formats (equivalent to 175 criterion-format records) was selected for validation against the ParER registry, which served as the human Gold Standard. With the aim of quantifying how close or similar the score given by the model is to that provided by the ParER register, the “Similarity %” metric was defined. This metric expresses the similarity between scores as a percentage from 0 to 100 (0% for completely different scores,

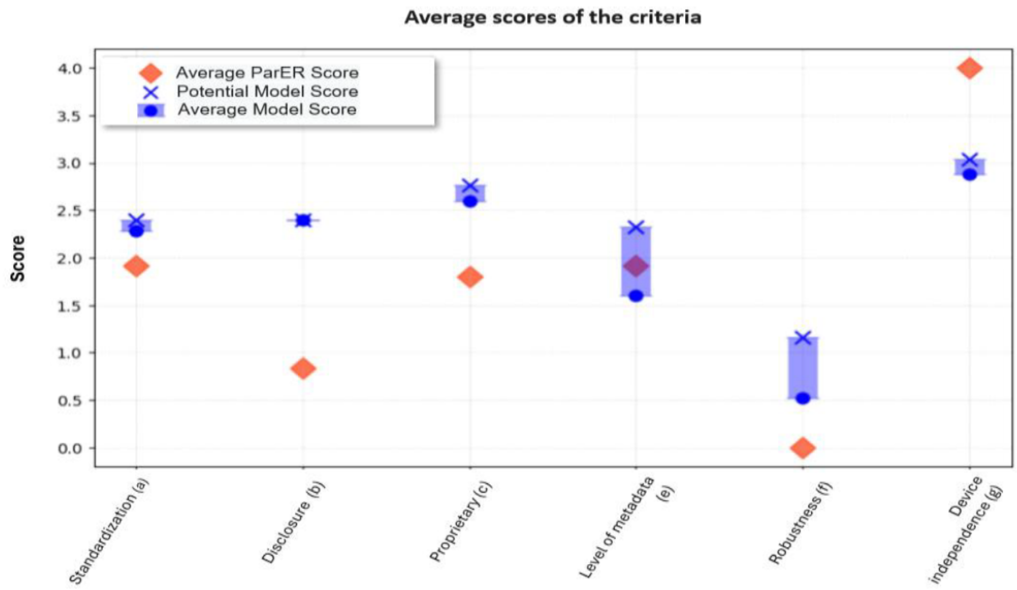

Figure 4: Average score of the criteria between the AI score and the ParER Gold Standard score.

100% for identical scores). This metric was formally defined with the following expression:

$$\text{Similarity \%} = \left(1 - \frac{\text{score difference}}{\text{maximum possible value for the criterion}}\right) \times 100 \quad (1)$$

To quantify the model's accuracy, two key metrics were employed. The first metric, "Average scores of the criteria" (Figure 4)[1], measured the percentage of times the AI score exactly matched the ParER Register score.

The chart compares the Average scores for each criterion across the 25 validation subset formats, showing results from both the AI model and the PARER registry. It also includes the model's average Potential score: the value the model would have achieved by replacing "I don't know" responses with the maximum possible class for that criterion. Based on this strict measure, the model achieved an exact coincidence in 58% of cases. Analyzing performance by criterion, the alignment was generally solid. In general, the similarity percentage achieves a high score, on average exceeding 65%. This indicates that, on average, even when the values do not coincide, the scores provided by the model are still close to those of the ParER register. Observing the graph, it can be deduced that the model's scores are, on average, close to those of the register for all six criteria considered. For the "Standardization" (a) and "Level of Metadata" (e) criteria there is a notable proximity between the model's average results and those of the ParER register. Factors related to "Device independence" (g) and "Robustness" (f) recorded high similarity rates of 72% and 74%, respectively. Conversely, the greatest challenge arose with the "Disclosure" (b) criterion, which registered the lowest similarity percentage, standing at 48%. This discrepancy was due to the model struggling to clearly distinguish between "open" and "closed" classes consistently with human evaluations, relying solely on the less categorical textual descriptions from the Library of Congress (LoC).

The second metric, "Interoperability Similarity %" measured the proximity of the scores as the difference between the values of the interoperability indices, compared to the maximum possible interoperability value of 21.

Figure 5 presents a graph displaying, for each of the 25 formats in the validation subset:

- Interoperability indices calculated by the model (represented by blue dots);
- Interoperability indices found in the PARER registry (represented by orange diamonds);

[1] In the chart shown in the figure, the Potential Model Score represents the value the model would have assigned if, instead of the "I don't know" class, it had assigned the class with the maximum value for that criterion.

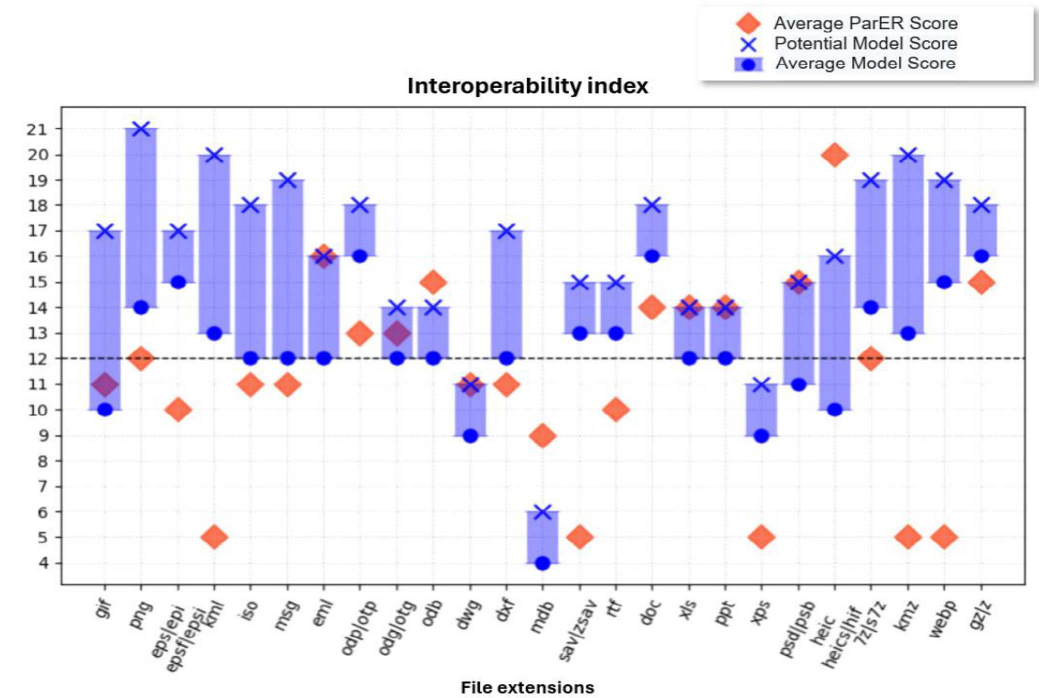

Figure 5: Interoperability Similarity % measures the proximity between the scores.

- The model's potential interoperability index, representing the value that could have been achieved if the model had assigned the maximum-value class for a given criterion instead of the "I don't know" classes (represented by blue x's).

The interoperability comparison across 25 file formats reveals a significant divergence between the two approaches: while ParER classifies 13 out of 25 formats as non-interoperable (52%), the AI model demonstrates greater flexibility, reducing this critical issue to only 6 out of 25 formats (24%). In terms of incremental efficiency, the AI identifies 9 formats as interoperable that ParER considers critical, whereas the opposite—interoperable for ParER but not for the AI—occurs in only 2 cases. On this more flexible basis, the system demonstrated high reliability, with a Similarity that exceeded 80% on the total index across most formats. This indicates that, while not perfectly replicating human judgment, the model provides a strong approximation. The model also showed a tendency to underestimate in certain cases, such as for the HEIC/HEICS/HIF and WEBP formats, which recorded the lowest Similarity (50%). Despite these specific discrepancies, the high similarity value confirms that the Artificial Intelligence system demonstrated the capacity to efficiently and traceably replicate the evaluations provided by the ParER Register (Figure 5). Analysing the model's performance, it is possible to highlight mainly two aspects:

- **Exact Coincidence**: The Average scores expressed by the AI model coincided exactly with those of the ParER Register in 58% of cases.
- **Total Similarity**: The "Interoperability Similarity %" on the total index was found to be high, exceeding 80% in most formats, demonstrating that the model's output is a strong approximation of human judgment.

Divergences emerged mainly when the conceptual frameworks of AgID and LoC differed significantly, forcing the model to interpret subtle nuances rather than relying on explicit statements.

The analysis of the scores highlighted that the model delivers results with a strong Similarity % compared to the ParER register, both for individual criteria and for the interoperability index. Although the Agid criteria have distinct and separate classes, the assessment of whether a LoC factor belongs to one class or another may be imprecise, due to descriptions of individual Agid criteria that are not always clear or easily distinguishable. This suggests that the model's evaluations could be improved by integrating more information into the Context parameter provided to the AI model.

A critical issue that emerged was the phenomenon of uncertainty. The evaluation class “I don’t know” was returned in 42 occurrences out of the total. It is important to note that the majority of these cases, exactly 25, were related to the “Extensibility” (d) criterion, which had been excluded from the mapping due to a lack of reliable data. Only 14 cases were classified as “I don’t know” due to insufficient or ambiguous information provided by the LoC source. It can be observed that cases classified as “I don’t know” may influence the final result, since they always receive a score of zero.

The “I don’t know” class provides greater reliability to the model, thanks to descriptions of criteria that are ambiguous or not very informative within the LoC. Even in this case, the model could further improve its assessments by introducing additional information to manage such cases more effectively.

These points of attention mainly suggest that to further enhance the model, it would be useful to provide more detailed and precise guidelines on the evaluation methods used for assigning the various classes. This approach will help refine the model, which already demonstrates strong assessment capabilities, making it even more accurate and reliable.

5. Conclusions and Future Work

This research has analyzed the complexity of periodic file format assessment within the AgID regulatory framework. In a digital ecosystem marked by rapid technological obsolescence, monitoring format sustainability represents a critical burden for digital curators. Consequently, the automation of this process emerges not merely as an operational necessity but as a required paradigm shift to ensure scalability and timeliness in long-term preservation.

Empirical analysis has revealed that the primary challenge lies in the conceptual misalignment between AgID criteria and the Library of Congress (LoC) sustainability factors. This divergence requires both human experts and Artificial Intelligence (AI) models to make interpretive leaps that may introduce subjectivity. However, the evolution of Large Language Models (LLMs) and the significant expansion of their context windows open unprecedented perspectives.

The ability to process massive volumes of data in a single pass allows for overcoming the rigidity of predefined mappings. Instead of limiting the AI to selected fields, future iterations could adopt a holistic approach, providing the model with the entire information content of LoC records (including textual notes, historical details, and extended technical specifications). This would allow for full utilization of the AI’s inferential reasoning, reducing the need for forced human mediation and improving assessment precision through a nuanced understanding of metadata and the integration of additional authoritative sources.

To consolidate the authority of the process, the results indicate the need to:

- Enrich frameworks with examples derived from real-world metadata;
- Integrate established technical registries (such as PRONOM or Wikidata) to create more robust preservation taxonomies;
- Evolve the prototype through fine-tuning, advanced prompt engineering, or hybrid approaches combining textual reasoning with structured metadata.

The project demonstrates that AI-assisted assessment holds full validity within digital preservation strategies. By reducing manual effort and ensuring transparent justifications, the system offers a scalable model for data sustainability. In alignment with FAIR principles (Findability, Accessibility, Interoperability, Reusability), alongside publishing its research results and making its work available for reuse, ParER is also considering developing a service to calculate the interoperability index and make it available to interested parties [26].

The prototype is not a mere exercise in automation, but proof of how archival institutions can successfully integrate cutting-edge computational tools to safeguard the digital heritage entrusted to their care.

Acknowledgments

We would like to thank all colleagues from the Emilia-Romagna Digital Archival Centre and the team at Engineering Group S.p.A. who collaborated on the project.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Digital Preservation Coalition, Digital Preservation Handbook – File formats and standards, 2025. URL: https://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards.
[2] ISO 13008:2022 Information and documentation – Digital records conversion and migration process, 2022.
[3] ISO 14721:2025 Space Data System Practices — Reference model for an open archival information system (OAIS), 2025.
[4] PRONOM, The online registry of technical information, 2025. URL: https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.
[5] Library of Congress, Sustainability of Digital Formats: Planning for Library of Congress Collections, . URL: https://www.loc.gov/preservation/digital/formats.
[6] Library of Congress, Library of Congress Recommended Formats Statement 2025–2026, . URL: https://www.loc.gov/preservation/resources/rfs.
[7] The National Archives of United-Kingdom, Selecting File Formats for Long-Term Preservation, . URL: https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.
[8] Digital Preservation Coalition, File Formats Assessments, . URL: https://wiki.dpconline.org/index.php?Title=File_Formats_Assessments.
[9] United States National Archives and Record Administration (NARA), Digital Preservation Risk Matrix, . URL: https://github.com/usnationalarchives/digital-preservation/tree/master/Digital_Preservation_Risk_Matrix.
[10] G. Drago, Recommended File Formats for Long-Term Archiving and for Web Dissemination in Phaidra, . URL: https://phaidra.cab.unipd.it/static/EN-file-formats.pdf.
[11] J. Rog, C. Van Wijk, Evaluating File Formats for Long-term Preservation, . URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.181.506&rep=rep1&type=pdf.
[12] Centre de coordination pour l'archivage à long terme de documents électroniques (CECO), Matrice d'évaluation, . URL: https://kost-ceco.ch/cms/evaluation.html.
[13] Recommended Preservation Formats for Electronic Records, . URL: https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records.
[14] Library and Archives Canada, File Format Guidelines for Preservation and Long-term Access, . URL: http://www.councilofnsarchives.ca/sites/default/files/LAC%20File%20Format%20Guidelines%20for%20Preservation%20and%20Long-term%20v1_2010-12_0.pdf.
[15] InterPARES 2, Selecting Digital File Formats for Long-Term Preservation. General Study 11. Final Report, . URL: http://www.interpares.org/ip2/ip2_case_studies.cfm?study=35.
[16] Agency for Digital Italy, Guidelines on the creation, management and preservation of electronic documents (Linee Guida sulla formazione, gestione e conservazione dei documenti informatici), 2021. URL: https://www.agid.gov.it/sites/agid/files/2024-05/linee_guida_sul_documento_informatico.pdf.
[17] Agency for Digital Italy, Annex 2 "File formats and migration" to the Guidelines on the creation, management and preservation of electronic documents (Linee Guida sulla formazione, gestione

e conservazione dei documenti informatici), 2021. URL: https://www.agid.gov.it/sites/agid/files/2024-06/allegato_2_formati_di_file_e_riversamento.pdf.

[18] S. Allegrezza, The Italian guidelines on creation, management and preservation of digital records. A proposed methodology for file formats assessment, in: Proceedings of iPRES 2021 17th International Conference on Digital Preservation, 2021, pp. 1–8. URL: https://www.ipres2021.ac.cn.

[19] ParER - Emilia-Romagna Digital Archive Centre, . URL: https://poloarchivistico.regione.emilia-romagna.it/english.

[20] Engineering Ingegneria Informatica Spa, GenAI supports archivists to preserve document accessibility over time, . URL: https://www.eng.it/en/insights/stories/case-studies/genai-al-servizio-dellarchivista-per-preservare-i-file-digitali-nel-tempo.

[21] ParER - Emilia-Romagna Digital Archive Centre, Giornata di studio "Archivi e intelligenza artificiale responsabile. Esperienze Opportunità Cautele", . URL: https://poloarchivistico.regione.emilia-romagna.it/polo-archivistico/eventi/giornata-di-studio-archivi-e-intelligenza-artificiale-responsabile.

[22] Formez, L'intelligenza artificiale negli archivi: soluzioni in campo e casi concreti, . URL: https://eventipa.formez.it/eventi/9fbaa3dd-0945-42d3-a842-1973822960f7.

[23] The National Archives, The technical registry PRONOM, . URL: https://www.nationalarchives.gov.uk/pronom.

[24] Wikidata, Dataformats, . URL: https://www.wikidata.org/wiki/Wikidata:Data_formats.

[25] Library of Congress, Mapping FDDs to PRONOM and Wikidata Unique Identifiers, . URL: https://www.loc.gov/preservation/digital/formats/fdd/fdd_puid_qid.shtml.

[26] GO FAIR Initiative, . URL: https://www.go-fair.org/fair-principles/.

Note: All URLs were last accessed December 11, 2025.