ARCHEION
2025, T. CXXVI, s. 283–289
ISSN 0066-6041
e-ISSN 2658-1264

DLM Forum Members' Meeting, Gdańsk, 4 June 2025

Rapid digitalisation of institutional processes presents both opportunities and risks for the archival sector. While e-services and big data create unprecedented possibilities for access, transparency, and reuse, they also raise significant challenges regarding sustainability, authenticity, and long-term preservation. This article examines four dimensions of archival transformation in the digital age: embedding archival foresight in the design of e-services, developing governance strategies for big data, leveraging artificial intelligence (AI) to enhance archival access, and fostering a culture of innovation within archival institutions. Drawing on the case of the European Parliament Archives, it illustrates how archival principles can be reconciled with emerging digital technologies to create a resilient and future-proof information ecosystem.

This approach was introduced during the DLM Forum Members' Meeting, Gdańsk (Poland), 4–5 June 2025.

Future-proofing archives in the digital age: integrating archival principles, big data management, and Artificial Intelligence

1. Introduction

Digital transformation of public administrations has profound implications for records and archives management. Institutions increasingly operate in an environment dominated by e-services, big data, and AI. While these tools promise efficiency and accessibility, they also, unless designed with sufficient foresight, pose a risk of undermining archival integrity. Without proactive strategies, archives face fragmentation, obsolescence, and loss of authenticity.

The archival discipline must therefore adapt its methodologies to ensure that principles such as provenance, authenticity, and accessibility remain relevant in digital contexts. This paper contributes to ongoing discussions on digital archival practice by analysing how institutional archives can integrate traditional archival principles with emerging technological paradigms.

2. Literature and context

Scholarly debates on digital archiving emphasise the importance of integrating archival principles into digital service design[1]. International standards such as ISO 15489[2] (Records Management) and MoReq2010 highlight the necessity of metadata, retention schedules, and authenticity verification in digital environments.

In parallel, the rise of big data has generated both opportunities for research[3] and challenges for governance, particularly with regard to scalability, privacy, and digital sovereignty. AI, and especially large language models (LLMs), is now emerging as a transformative force in archival access[4].

The European Parliament Archives provide a unique case study of how these challenges are confronted within a large multilingual institution, operating under strict legal and democratic accountability requirements.

---

1 L. Duranti, *Concepts and principles for the management of electronic records, or records management theory is archival diplomatics*, "Records Management Journal" 2010, vol. 20(1), pp. 78–95; P. Conway, *Preserving imperfection: Assessing the authenticity of digital surrogates*, "Journal of the American Society for Information Science and Technology" 2015, vol. 66(12), pp. 2576–2589.

2 ISO 15489-1:2016 *Information and documentation – Records management – Part 1: Concepts and principles.*

3 R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, Thousand Oaks 2014.

4 M. Buckland, *Information and Society*, Cambridge, Massachusetts 2021, https://doi.org/10.7551/mitpress/10922.001.0001; E. Frontoni, M. Paolanti, T.P. Lauriault, M. Stiber, L. Duranti, M. Abdul-Mageed, *Trusted Data Forever: Is AI the Answer?* [in:] *Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29 – April 1, 2022), Edinburgh, UK,* https://arxiv.org/pdf/2203.03712 [access: 3.10.2025].

3. Methods and approach

This article adopts a case study approach, examining the strategies and innovations implemented within the Archives Unit of the European Parliament. It analyses policy frameworks, system architectures, and applied technologies to illustrate how archival foresight, big data governance, and AI-enhanced access can be combined in practice.

Four dimensions guide the analysis:

1. Design of e-services embedding archival principles.
2. Governance of big data through structured workflows and AI-driven curation.
3. Accessibility strategies leveraging natural language processing and retrieval-augmented generation.
4. Institutional innovation culture, emphasising collaboration across disciplines.

3.1. Designing e-services with archival foresight

E-services have revolutionised institutional interactions with citizens, yet they often neglect long-term records management. Integration of archival considerations at the design stage is essential to avoid fragmented and inaccessible digital legacies. Key requirements include application of metadata standards, use of sustainable storage formats (e.g., PDF/A for written documents), digital signatures for authenticity verification, and systematic application of retention schedules.

In the European Parliament, these principles have been operationalised through a unified Archive Management System for written documents, a separate system for multimedia, and the development of a central Records Management System. Despite this progress, existence of a huge number of heterogeneous IT solutions illustrates the complexity of achieving harmonisation across institutional infrastructures.

3.2. Big data: from opportunity to liability

Proliferation of digital information in the form of "big data" raises pressing questions about governance, preservation, and access. While large-scale datasets provide new opportunities for research and policy analysis, they also risk becoming an overwhelming liability if left unmanaged. Archival strategies must therefore evolve to handle the velocity, variety, and volume of data in contemporary institutions.

This requires robust technical infrastructures, scalable indexing systems, and AI-driven curation methods that facilitate real-time classification and filtering. Ethical considerations – such as data protection, user consent, and digital sovereignty – are equally critical. Within the European Parliament, practical steps include transition from paper-based to digital-first workflows, creation of a digital repository containing over 2.3 million digital documents in 2024, and deployment of AI tools for both automatic indexing (using EUROVOC) and generative pre-processing of fonds.

3.3. Enhancing access through Artificial Intelligence

Preserving digital data is only meaningful if it remains accessible and intelligible over time. Traditional archival description methods are insufficient for datasets of institutional scale. Emerging techniques in linked data, natural language processing, and AI-based classification hold the potential to transform discoverability and usability.

The Archives Unit of the European Parliament has pioneered innovative applications of AI[5], most notably through the development of Archibot[6]. This large language model (LLM)-based tool enables citizens and researchers to query archival holdings in natural language. By combining keyword-based search, semantic vector similarity retrieval, and re-ranking algorithms, the system ensures precise contextual results. Importantly, its multilingual design allows

[5] Historical Archives European Parliament, Archives Unit Dashboard, https://historicalarchives.europarl.europa.eu/home/cultural-heritage-collections/news/dashboard-tutorial.html [access: 3.10.2025]

[6] European Parliament, EP Archives Overview Dashboard, https://archidash.europarl.europa.eu/ep-archives-anonymous-dashboard [access: 3.10.2025].

queries in all 24 official EU languages, even when documents exist only in their original linguistic form.

The implementation of Retrieval-Augmented Generation (RAG) ensures that responses are grounded in archival sources, avoiding the hallucination problem often associated with generative AI. In this respect, AI does not replace archivists' expertise but complements it, offering new ways to explore and interpret institutional memory.

3.4. Fostering a culture of innovation in archives

The digital transformation of archives requires more than technology adoption: it requires an institutional culture that values experimentation and interdisciplinary collaboration. Archivists must work alongside IT developers, legal advisors, and data scientists to ensure that digital infrastructures respect archival principles while remaining adaptable to technological change.

Open-source technologies, low-code platforms, and community-driven solutions allow archives to tailor systems to their specific mandates. By implementing a solution with generative AI applied at the fonds level, rather than at the level of individual documents, the Archives Unit demonstrates how archival methodologies can be translated into innovative digital services.

Let's take the example of our generative AI solution. If data is to be the fuel for AI, it must first be properly prepared for AI algorithms to use. In our system, we don't handle individual documents – we work with fonds. A fonds is a collection of documents that naturally originate from the activities of a specific agency, person, or organisation.

Once a fonds is processed by archivists, we upload all its documents along with their metadata. Metadata includes details like the document type, reference number, file name, title, and creation year.

Next, we divide each document into overlapping chunks of about 300 tokens (words and characters). Each chunk is converted into an embedding, which is a numerical representation of the text. We use an Embedded Multilingual model to create a 1024-dimensional vector for each chunk.

Finally, we store and index the metadata, the text chunk, and its embedding in our database using OpenSearch.

At the heart of our AI solution is a powerful RAG. First, we retrieve only the relevant European Parliament documents for a user query. A context retriever is responsible for selecting relevant information from the collection of documents proposed in response to this query after applying guardrails.

The context retriever integrates keyword-based search (named-entities to perform a keyword-based BM25 search using OpenSearch) and with a vector similarity search (k-NN vector search based on a cosine similarity using the embedded model). The 10 best results are retrieved from the keyword-based search. The 20 best documents are retrieved from the semantic search. These 30 best documents are merged and followed by a re-ranking stage to optimise the relevance of retrieved chunks. The maximum 8 best documents with a reranking rate superior to 0.4 of the highest score are kept.

Once this context is established, the AI generates answers based on this information to answer the query and considering the hypertext parameters, synthesizing information from these multiple sources. The solution's prompt is configured to be responsible when generating the answer using a broad language model that is constitutional AI, setting the creativity value of the system to 0, the temperature.

But what makes our system truly unique is its ability to work in many languages. The European Parliament operates in 24 official languages, but many historical documents exist only in their original language. Using multilingual search capabilities, our AI solution can find and summarize chunks, even if they have never been translated into the language of your query. Whether you ask questions in Maltese about a debate in 1965 or about a policy in Norwegian in 1992, you will receive accurate and contextual answers in the language of your choice.

4. Conclusion

The digital age presents the archival sector with both existential risks and transformative opportunities. If archives are to remain relevant, they must embed foresight into e-service design, manage big data through structured governance, harness AI for enhanced accessibility, and cultivate a spirit of innovation.

The case of the European Parliament Archives illustrates that it is possible to reconcile traditional archival principles with emerging technologies, thereby

creating a resilient and democratic information infrastructure. Future research should further investigate the ethical governance of AI in archives, cross-institutional metadata interoperability, and citizen-oriented archival interfaces.

Ludovic Delépine

European Parliament (Luxembourg)
ludovic.delepine@europarl.europa.eu