ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

Scientific Communication and Augmented Intelligence in the Datafication of the Documentary Record

Heber Tormentino de Sousa

Federal University of São João del-Rei, Brazil

Vinícius Braga

Federal University of Minas Gerais, Brazil

Gercina Ângela de Lima

Federal University of Minas Gerais, Brazil

Abstract

This essay analyzes the Datafication of the Documentary Record (DDR) in scientific communication, a process that converts articles into algorithmic processing units, resulting in critical losses of narrativity, provenance, and authorial context. From the perspective of Knowledge Organization, the text warns that this atomized fragmentation may promote the erasure of the scientific dialectic and reinforce power asymmetries typical of data colonialism. As a solution, it proposes the paradigm of Augmented Intelligence (AugI), which advocates systematic human oversight to restore the document's layers of meaning. In this scenario, the figure of the 'Augmented Curator' emerges, referring to mediators who utilize technology to expand their critical capacity without relinquishing ethical responsibility, ensuring that the scientific record remains a space for human dialogue, integrity, and veracity.

1.0 Introduction

The transformation of dimensions of human experience into data suitable for computational processing, datafication, has been the subject of increasingly sophisticated analyses. From an incisive critical perspective, Couldry and Mejias (2018) propose understanding massive data extraction as a form of colonialism, in which the appropriation of daily life, converted into informational flows, establishes new power asymmetries. It is within this context that the Datafication of the Documentary Record (DDR) in scholarly communication is situated, a phenomenon investigated in this essay.

Knowledge Organization (KO), a field within Library and Information Science (LIS), is concerned with the technical representation of knowledge artifacts (articles, books, and journals) to promote efficient information retrieval. To this end, it is guided by fundamental ethical principles such as neutrality, provenance, and the non-dogmatic character of science (Fox and Reece 2012; Tognoli and Guimarães 2019; Popper 1934).

Documents, especially the scientific documentary record, the object of this reflection, are composed of medium, formatting, and text. However, the processing carried out by massive Knowledge Technologies (Schreiber et al. 2000) retains only the text, dismantling the original narrative into data.

The process of text atomization, referred to here as the Datafication of the Documentary Record (DDR), converts the document into digital raw material, alienating it from its medium and formatting. By shifting the unit of meaning from text to data, DDR inaugurates new forms of organization that bring epistemological and ethical challenges to information seeking, demanding well-informed criticism (Gitelman 2013).

This reflection seeks to answer the following question: what does a scientific article become when it is no longer intended for human reading and begins to be produced for automated processing? To answer it, a Narrative Literature Review is adopted, constructing a theoretical argument based on seminal authors from KO and related fields. The discussion covers the ontology of the document, power within classification systems, the challenges of DDR, and the potential of augmented intelligence as a paradigm for reconfiguring scholarly communication.

It is argued that while DDR increases productivity, it entails ethical risks that can be mitigated through mediation based on planned collaboration between humans and automated systems, termed Augmented Intelligence (Engelbart 1962).

2.0 Methodology

This theoretical essay, reflective in nature, is grounded in bibliographic research (Gil 2017) to critically analyze the tensions between document and data in scholarly communication considering the Datafication of the Documentary Record (DDR). The literature selection explored conceptual frameworks that support the thesis of Augmented Intelligence as an ethical-normative paradigm for scientific curation within the context of massive knowledge technologies. This work does not intend to provide a definitive analysis, but rather to offer a dialogic contribution open to new perspectives.

3.0 Ontological Foundations

This section establishes the conceptual pillars: the document as an ontological entity, datafication as a technical phenomenon, and Augmented Intelligence as an ethical-normative paradigm. Following Briet (2006), the document is evidence or cultural trace, a medium that carries information and functions as material evidence of situated human activity. Within the scientific sphere, it possesses narrative integrity and authorial context, being an inseparable unit from its historical and social dimensions. It is important to emphasize that the term "cultural trace" is a very common contemporary interpretation in LIS (Library and Information Science), though it is not Briet's (2006) strict terminology.

Datafication, as defined by Mayer-Schönberger and Cukier (2013, 78), is the act of quantifying phenomena for tabulation and analysis. In DDR, the document's continuous narrative is fragmented into its operational units and reorganized into numerical vectors (embeddings), which allow for semantic proximity calculations without the grasp of meaning. Unlike digitization, which merely converts the medium, DDR undergoes a profound transformation: the document shifts from letter to number, from cultural trace to a resource for algorithmic processing.

This reconfiguration imposes ethical dilemmas, such as access to research data. Borgerud (2025) warns that access is not an absolute value, requiring a balance between democratic interests and ethical safeguards. Thus, the need for critical mediation is reinforced.

It is in this context that Augmented Intelligence (AugI) emerges as a necessary paradigm. Originally proposed by Engelbart (1962) as the enhancement of human cognitive capacity through technology, Augmented Intelligence is redefined here, following Sousa and Lima (2025), as the systematic inclusion of human oversight in the automated processes of Knowledge Organization. Under this ethical-normative paradigm, technology does not replace humans; on the contrary, human discernment, integrated into the DDR workflow, can mitigate the losses of provenance, context, and meaning that datafication tends to provoke. Thus, this ontological foundation describes the phenomenon while pointing the way toward critical mediation.

4.0 Information Loss in DDR

The conversion of a document into computational input is not a neutral operation, as it imposes significant losses that affect the integrity of scholarly communication. By applying granular processing to the original narrative to make it processable, the Datafication of the Documentary Record (DDR) sacrifices constitutive dimensions of the document as a cultural trace.

4.1 Loss of Narrativity

The document-as-cultural-trace possesses an internal organization (titles, sections, graphical emphases, visual hierarchy) that guides human interpretation. Since only the text is extracted in DDR, visual semantics are lost. Elements such as paragraph positioning, the use of italics for foreign terms, or the hierarchy of sections are flattened in the tokenization process.

4.2 Loss of Provenance and Context

By transforming text into embeddings during the application of DDR, semantic proximity relations are captured without considering the limitations, assumptions, and production context of the article (Merton 1942). In vector representation, statements are stripped of

their authorship and their conditions of validation, circulating instead as autonomous entities.

As Gitelman (2013) argues, data is always "cooked", that is, generated to meet specific demands. The resulting data do not carry the caution inherent to scientific authority, nor the constraints imposed by the historical context of production. Thus, ideas, perceptions, and statements are reduced to numerical vectors to facilitate algorithmic efficiency, at the expense of the epistemological precision that Knowledge Organization seeks, by principle, to preserve. However, it should be noted that from this perspective, DDR is a second layer of fragmentation, as the author of the document, when writing the scientific article, has already performed their own "cooking" of reality.

5.0 The Ontological Transformation of the Document-as-Index

While the previous section analyzed losses in their concrete manifestations (medium, provenance, narrativity), it is now necessary to investigate how these losses, taken together, perform an ontological transformation of the document-as-index, as its very nature is altered when subjected to DDR.

The dissolution of the scientific article as an "entity" in DDR occurs when the narrative unit is replaced by an ontology of things, composed of atomic fragments (tokens) that circulate detached from their origin. This new ontology is not that of the author's discourse, but that of computational representation. The document loses its autonomy as an individual voice within the scientific ecosystem, becoming a dispersed set of decomposed parts, processed and indexed by algorithms.

One of the most significant losses imposed by the Datafication of the Documentary Record (DDR) occurs at the level of representing veracity and scientific assertion. Upon extracting information from a document, the system faces a fundamental ontological choice, often implicit in its own architecture. We propose a distinction between two approaches (Table 1):

1. Strong Ontological Commitment (Representation as Fact): The system extracts information and records it as a domain truth. For example: "Particle X causes reaction Y." The data is alienated from its origin and begins to exist as an autonomous statement in the database, stripped of authorship, context, and the caveats provided by the original author.
2. Weak Ontological Commitment (Representation as Statement): The system records that "Author A, in article B, asserts that particle X causes reaction Y." This preserves the discursive character of knowledge and the provenance of the assertion.

Table 1. Comparison between Strong Ontological Commitment (representation as fact) and Weak Ontological Commitment (representation as statement) in the context of DDR

Characteristic	Strong Ontological Commitment (Fact)	Weak Ontological Commitment (Statement)
Representation	Extracts information as an absolute domain truth.	Records that "Author A asserts X" within a specific context.
Example	"Particle X causes reaction Y."	"In article B, it is asserted that particle X causes reaction Y."
Provenance	Alienated from its origin; autonomous statement stripped of authorship.	Preserves the discursive character and the source of the assertion.
Risk/Benefit	Risk of erasing scientific dialectics and promoting dogmatism.	Maintains doubt, debate, and knowledge traceability.

The document-as-trace, by its nature, carries doubt, debate, and provisional character that define science, including within documentary anabiosis (Couzinet 2016). From a technical standpoint, the dogmatism perceived by users in LLMs, which operate through DDR, is an effect of the text output, as they internally process neural weights. By favoring a strong ontological commitment, whether for economic reasons, efficiency, or simplicity, one runs the risk of promoting the erasure of scientific dialectics. Scientific knowledge, theoretically a peer-to-peer dialogue mediated by documents, becomes a probabilistic network of data stripped of its authorial voice.

Considering this scenario, proposals such as the publication fact labels by Willinsky and Pimentel (2024) counteract the erasure promoted by DDR. By suggesting that scientific articles be accompanied by a public and professional guide that explicitly details editorial decisions, review processes, and the limitations of findings, the authors offer a concrete mechanism to restore the layer of provenance and context that strong ontological commitment tends to suppress. It is a device that reinserts doubt and dialectics into the flow of scholarly communication, even when mediated by machines.

Computational heuristics prioritize a statistically plausible response but lack the authority and prudence of the original author. Consequently, under a strong commitment, authorship and responsibility for the statement become irrelevant. The document-as-index is reduced to a mathematical representation, unlike traditional Knowledge Organization Systems (KOS), which maintain an explicit link between the concept and its source.

The reduction of the document to a mathematical representation is a design choice. Stocker et al. (2025) propose that research findings be published from the outset as machine-readable expressions, maintaining precision and traceability, anticipating computational readability, and ensuring a weak ontological commitment. This critical

integration from the genesis [of the record] offers an alternative to the predatory logic of DDR.

6.0 From Raw Data to Invisibility

The culmination of datafication lies in the belief that data extracted from scientific articles constitute an objective and neutral representation of knowledge. However, when tracing the document's trajectory toward its reduction into statistical vectors, it becomes evident that the concept of "raw data" is, itself, a constructional discursive layer that conceals the decisions, selections, and subjectivities that permeate the entire process.

DDR promotes an invisibility that is inversely proportional to its technical efficiency, and this invisibility has a structural effect. This "invisibility" presents itself as a neutral layer when, in fact, it carries the bias of those who designed the system, those who defined what is relevant, and those who established the rules for connecting the fragments.

The naturalization of data as a neutral processing component is integral to the logic of informational capitalism. Couldry and Mejias (2018) define this process as data colonialism, where the conversion of daily life into information flows normalizes human exploitation, analogous to the appropriation of territories in historical colonialism. These data relations establish a social order of continuous tracking, enabling new forms of discrimination and behavioral control.

In scholarly communication, this dynamic transmutes the documentary record into raw material for technological infrastructures that, under the guise of efficiency, reinforce power asymmetries. Thus, DDR ceases to be a mere technical process and becomes a capture device that operates the "capitalization of life without limit" (Couldry and Mejias 2018, 337).

The epistemological danger lies precisely in accepting the output of these systems as the ultimate truth of science, ignoring that "data" is a fragment of a cultural trace, synthesized to fit the rigid structures of Large Language Models (LLMs). Interrogating raw data within DDR constitutes, in this sense, an act of epistemological resistance for Knowledge Organization.

If data is a construction, it can and must be audited; if datafication involves selections and "cuts," these must be made explicit and subjected to critical scrutiny. The passive acceptance of mass DDR leads to a science based on statistical correlations at the expense of causality and the deep understanding that characterizes scientific knowledge.

At this point, the critique of the invisibility produced by DDR finds a response, as outlined in the next section, which points to the need to place the human back at the center of the process, not as a mere machine operator, but as an auditor capable of restoring the sacrificed layers of meaning.

7.0 Augmented Intelligence and the Emergence of the Augmented Curator

As a response to the risks of DDR, Augmented Intelligence (AugI) proposes augmented curation: technology enhances the discernment of the information professional, positioning them as a critical mediator between the original document and its datafied representation, unlike LLMs, which seek to replace human cognition.

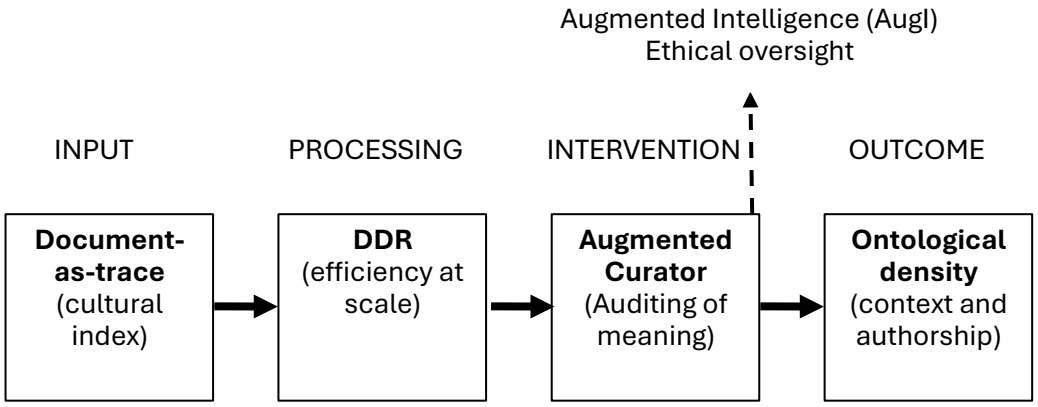


Figure 1. The Augmented Curation Workflow: integrating algorithmic efficiency (DDR) with human meaning-auditing to preserve ontological density within the scientific record

Within the framework of Augmented Intelligence (Figure 1), artifacts for intellect augmentation are investigated through systems integrated into the cognitive workflow that expand memory and decision-making. The Augmented Curator, from this perspective, interacts with these artifacts to enhance their critical judgment, acting as an auditor of meaning who corrects distortions such as LLM hallucinations and algorithmic biases, ensuring that DDR does not silence the multiplicity of scientific voices (Engelbart 1962; Xia and Maes 2013).

Regarding system architecture, Augmented Intelligence, therefore, demands technologies designed for weak ontological commitment, reducing assertions to facts only within their original context. This implies designing Knowledge Organization Systems (KOS) that make evidence and sources explicit, placing computational efficiency at the service of traceability. In this direction, Vogt (2025) proposes the CLEAR principle, which places the human back at the center by ensuring cognitively interoperable digital objects, reconciling automation with semantic density.

Thus, human mediation is not optional but rather the pillar that restores the document's ontological density, such as authorship, context, and provenance, which datafication erases. An integration is proposed where the machine processes at scale while maintaining the curator in spaces where auditing meaning is both necessary and humanly processable. In this sense, Augmented Intelligence establishes itself as an ethical-normative paradigm,

ensuring that scholarly communication preserves its social function of emancipation and the integrity of meaning.

8.0 Concluding Remarks

This essay has proposed an initial path of reflection on the Datafication of the Documentary Record (DDR) in scholarly communication, without claiming to be exhaustive. The challenge ahead is precisely to ensure that, even when enhanced by technologies, intelligence remains essentially human, ethical, and committed to preserving knowledge in its full narrative plenitude.

Traversing the transition from document-as-trace to document-as-data, significant risks to the fundamental principles of Knowledge Organization (KO) were identified, such as the loss of provenance, the invisibility of peripheral productions, and the suppression of scientific dialectics. These stem from the technical opacity of massive processing systems, particularly Large Language Models (LLMs). It has become evident that fragmenting text into data creates a device that distances the information mediator from the original narrative; however, the answer lies not in *a priori* rejection of technology, but in the critical reconfiguration of its use.

It was in this context that Augmented Intelligence (AugI) presented itself as an ethical-normative paradigm for augmented curation, placing machine efficiency at the service of human discernment and expanding the reach of well-informed decisions without replacing the mediator or the capacity for critique. From this perspective, massive knowledge technologies give rise to the figure of the Augmented Curator to mitigate the effects of DDR.

The limitations of this investigation are acknowledged, considering that AugI applies to humanly processable contexts where human mediation is strictly necessary, such as research institutions. What has been presented here is only one of many possible routes, and the essay-like nature of this work is an invitation to open debate. The future of scholarly communication will depend on our collective ability to investigate and contest digital infrastructures with the same rigor historically applied to physical media.

The transition from document-as-trace to document-as-data should not be rejected but critically reconfigured. The future of scholarly communication requires Knowledge Organization to remain essentially human to ensure informational justice, integrity, veracity, and the dialectic of meaning within the scientific record.

Artificial intelligence usage statement: The authors inform that the DeepSeek-V4-Flash model, accessed via the web interface in fast mode, was employed for the translation of the manuscript from Brazilian Portuguese into the official language of the event.

References

Borgerud, Christine. 2025. “The Public Right of Access to Research Data – An Ethical Dilemma?” Journal of Documentation 81, no. 7: 355–75.

Briet, Suzanne. 2006. What Is Documentation? English Translation of the Classic French Text. Translated and edited by Ronald E. Day and Laurent Martinet. Lanham, MD: Scarecrow Press.

Couldry, Nick, and Ulises A. Mejias. 2018. “Data Colonialism: Rethinking Big Data’s Relation to the Contemporary User.” Television & New Media 20, no. 4: 336–49. https://doi.org/10.1177/1527476418796632.

Couzinet, Viviane. 2016. “A Documentologic Approach of Herbarium: Documentary Anabiosis and Philogenetic Classification.” Proceedings from the Document Academy 2, no. 1.

Engelbart, Douglas C. 1962. Augmenting Human Intellect: A Conceptual Framework. Summary Report AFOSR-3223. Menlo Park, CA: Stanford Research Institute.

Fox, Melodie J., and Austin Reece. 2012. “Which Ethics? Whose Morality?: An Analysis of Ethical Standards for Information Organization.” Knowledge Organization 39, no. 5: 377–83.

Gil, Antonio Carlos. 2017. Como elaborar projetos de pesquisa. 6th ed. São Paulo: Atlas.

Gitelman, Lisa, ed. 2013. “Raw Data” Is an Oxymoron. Cambridge, MA: MIT Press.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston: Houghton Mifflin Harcourt.

Merton, Robert K. 1942. “A Note on Science and Democracy.” Journal of Legal and Political Sociology 1: 115–126.

Popper, Karl. 1934. A lógica da pesquisa científica. Translated by Leonidas Hegenberg and Octanny Silveira da Mota. São Paulo: Cultrix.

Schreiber, Guus, Hans Akkermans, Anjo Anjewierden, Robert de Hoog, Nigel R. Shadbolt, Walter Van de Velde, and Bob J. Wielinga. 2001. Knowledge Engineering and Management: The CommonKADS Methodology. 2nd ed. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/4073.001.0001

Sousa, H. T., and G. A. Lima. 2025. “Inteligência aumentada: Um caminho ético para a prática profissional na Biblioteconomia e Ciência da Informação.” Em Questão 31: 1–31.

Stocker, Markus, Lauren Snyder, Matthew Anfuso, Oliver Ludwig, Freya Thießen, Kheir Eddine Farfar, Muhammad Haris, Allard Oelen, and Mohamad Yaser Jaradeh. 2025. “Rethinking the Production and Publication of Machine-Readable Expressions of Research Findings.” Scientific Data 12: 677. https://doi.org/10.1038/s41597-025-04905-0.

Tognoli, Natalia Bolfarini, and José Augusto Chaves Guimarães. 2019. “Provenance as a Knowledge Organization Principle.” Knowledge Organization 46, no. 7: 507–15.

Vogt, Lars. 2025. “The CLEAR Principle: Organizing Data and Metadata into Semantically Meaningful Types of FAIR Digital Objects to Increase Their Human Explorability and Cognitive Interoperability.” Journal of Biomedical Semantics 16 (18).

Willinsky, John, and Daniel Pimentel. 2024. “The Publication Facts Label: A Public and Professional Guide for Research Articles.” Learned Publishing 37, no. 2: e1599.

Xia, Chao-Ti, and Pattie Maes. 2013. “The Design of Artifacts for Augmenting Intellect.” In Proceedings of the 4th Augmented Human International Conference, 154–61. New York: ACM.