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

Metadata Curation and Data Quality in Cultural Heritage Repositories: The Case of the Museu Paulista Collections

Rodrigo Freire de Oliveira

University of Brasília, Brasília, DF, Brazil

Tatiana Vasconcelos dos Santos

Museu Paulista, University of São Paulo, São Paulo, SP, Brazil

Giovanna Fulan Augusto Balsan

Museu Paulista, University of São Paulo, São Paulo, SP, Brazil

Solange Ferraz de Lima

Museu Paulista, University of São Paulo, São Paulo, SP, Brazil

Abstract

Since the 1990s, the Museu Paulista of the University of São Paulo has carried out initiatives aimed at organizing, documenting, and digitally disseminating its collections. Following the migration of its documentary database, originally developed in Visual FoxPro, to Tainacan, a process was initiated to review and enhance the quality of the metadata of the Three-Dimensional and Iconographic collections. This applied and qualitative article analyzes this process based on technical reports, diagnostic spreadsheets, and operational records produced during the project. The discussion brings together Knowledge Organization, data quality, metadata quality, and information retrieval, drawing on authors such as Hjørland, Guimarães, Wang and Strong, Pipino, Lee and Wang, Bruce and Hillmann, Park, and Ochoa and Duval. The results show a reduction in terminological dispersion, the conceptual restructuring of metadata, the creation of institutional criteria for field completion, and the consolidation of information curation practices. The article concludes that the treatment analyzed is not limited to the syntactic cleaning of data, but constitutes an applied practice of Knowledge Organization in a museological context.

1.0 Introduction

Since the early 1990s, the Museu Paulista of the University of São Paulo has carried out systematic initiatives to organize, document, and digitally disseminate its collections. During this period, the foundations of the Museu Paulista Documentary System[1] were established. Developed within the institution itself, the system was designed to record and organize information related to the Three-Dimensional and Iconographic collections. The database, initially built in Visual FoxPro, was populated over approximately three decades by different teams, projects, fellows, researchers, and professionals working in museological documentation.

Between 2020 and 2022, after assessing the functionalities of Tainacan, an open-source platform widely adopted in the Brazilian museological field, the Museu Paulista conducted the migration of data from its local database to this new online digital infrastructure. The migration expanded the possibilities for public dissemination of the collection, but it also made visible problems accumulated over the course of the database's internal use, such as terminological variation, inconsistencies in field completion, the presence of non-standardized multi-valued entries, the use of abbreviations, fields with inappropriate data types, and information gaps.

This scenario is common in cultural heritage institutions where collection documentation develops over time in articulation with research, teaching, outreach, curation, training, and institutional management. Museological databases are not static products, but the result of institutional policies that affect the practices through which objects are represented, described, interpreted, and reviewed. For this reason, improving information quality cannot be reduced to the technical cleaning of spreadsheets: it involves conceptual decisions analyzed in light of a historical understanding of the institution's trajectory and of how cultural objects are represented, retrieved, and communicated.

This article analyzes the metadata curation process developed at the Museu Paulista after the migration to Tainacan. Its objective is to examine how diagnostic procedures, level-of-effort classification, technical treatment, curatorial validation, and semantic reorganization contribute to improving metadata quality in a cultural heritage digital repository. The study is guided by the following question: how can metadata curation, articulated with Knowledge Organization and data quality, enhance the representation, retrieval, and interoperability of information in museological repositories?

2.0 Literature Review

Metadata curation in cultural heritage digital repositories brings together Knowledge Organization, data quality, metadata quality, and information retrieval. In museological institutions, metadata are not merely technical elements of description, but structures of

1 The Museu Paulista Documentary System was developed and patented by Luciano Beraldo, an IT staff member at the institution, between 1993 and 1995, with further improvements in the following decades.

representation that shape the identification, classification, retrieval, interpretation, and reuse of cultural objects. The treatment of metadata in legacy databases should therefore be understood not only as the syntactic cleaning of data, but as the conceptual reorganization of information.

In the field of Knowledge Organization, Hjørland (2008) distinguishes knowledge organization processes, such as classification, indexing, description, and cataloging, from knowledge organization systems, such as taxonomies, thesauri, authority lists, and controlled vocabularies. This distinction is relevant because it shows that the quality of representation depends both on the procedures applied to data and on the terminological structures that organize a collection. Guimarães (2008), in the context of ISKO, reinforces the interdisciplinary character of Knowledge Organization and its mediating role among documents, systems, and users.

The domain analysis proposed by Hjørland (2017) helps to show that the terms used in museological collections are not neutral units, but express documentary practices, curatorial interpretations, and institutional conventions. For this reason, metadata normalization requires specialized knowledge: graphically similar terms may refer to distinct entities, while different terms may represent the same concept. In this sense, metadata curation involves both technical and interpretative decisions.

Information organization is also related to cataloging and metadata production. Novellino (1996) understands information representation as the selection and formalization of attributes capable of mediating the relationship among documents, systems, and users. Mey (1995) emphasizes that cataloging produces encoded messages that make documents identifiable and retrievable. In the digital environment, this function is operationalized through metadata, whose review involves not only the values entered into fields, but also the adequacy of field data types, the consistency of vocabularies, and the repository's capacity for interoperability. Gilliland (2016) reinforces that metadata are grounded in standards of structure, value, content, and communication, all of which are essential dimensions for interoperable digital environments.

The literature on data quality broadens this discussion by arguing that quality cannot be reduced to accuracy. For Wang and Strong (1996), high-quality data are data that are fit for use. This perspective is especially relevant to cultural heritage repositories, where a piece of information may be historically correct but recorded in an inappropriate field, thereby impairing its interpretation or retrieval. Ballou et al. (1998), Pipino et al. (2002), Bruce and Hillmann (2004), and Park (2009) highlight dimensions such as accuracy, consistency, completeness, timeliness, coherence, accessibility, and conformity to expectations, which are fundamental to data and metadata quality.

Indicators such as empty fields, number of unique terms, duplicates, inconsistencies in field completion, and terminological dispersion can be measured objectively. However, deciding whether a term is semantically appropriate requires contextual assessment by specialists. Tools such as OpenRefine support this process by enabling the identification of patterns, textual transformations, the grouping of variants, and the reduction of informational noise. As Ochoa and Duval (2009) observe, automatic procedures can

support the assessment of metadata quality, but they do not replace human judgment in semantic decisions.

Data and metadata quality have a direct impact on information retrieval. Saracevic (1999) situates information retrieval as a central concern of Information Science, oriented toward mediation among systems, documents, and users. Ingwersen and Järvelin (2005) expand this perspective by integrating information seeking, retrieval, and context. In cultural repositories, metadata must be technically consistent in order to enable search, filters, and interoperability, while also remaining intelligible to different audiences, including researchers, documentation professionals, curators, educators, and visitors. Metadata curation can therefore be understood as an applied practice of Knowledge Organization aimed at expanding the semantic consistency, retrievability, and interoperability of digital collections. This form of data curation thus operates on the two fronts distinguished by Hjørland (2008): knowledge organization processes and knowledge organization systems. It also occupies a central position in data governance within an institution dedicated to the preservation of cultural objects, insofar as it should not only address workflows and procedures, but also ensure that information representation expresses the institution's policy.

3.0 Methodology

The research is applied in nature and adopts a qualitative approach, supported by quantitative indicators. It takes the form of a case study centered on the metadata curation process of the Three-Dimensional and Iconographic collections of the Museu Paulista of the University of São Paulo. The documentary corpus analyzed consists of technical reports, action plans, diagnostic spreadsheets, meeting records, and operational logs produced during the review and treatment of the database.

The methodology was organized into six stages: diagnostic assessment of the database; classification of metadata by level of effort and definition of a treatment strategy; implementation of technical procedures for data normalization and transformation; semantic validation of the results; conceptual reorganization of specific metadata elements; implementation in the database; and documentation of the decision-making process.

In the diagnostic stage, quantitative and qualitative characteristics of the metadata were surveyed, including the number of completed items, number of unique terms, empty fields, special characters, occurrence of multiple values, inappropriate data types, and need for semantic restructuring. Based on this diagnosis, the metadata were classified into low, medium, and high levels of effort, considering semantic dispersion, consistency, and complexity. This classification made it possible to distinguish metadata that could undergo standardized treatment from those requiring contextual analysis and specialized validation.

The technical treatment was carried out with the support of OpenRefine, which was used both to preserve the traceability of changes and to perform data cleaning, transformation, and normalization procedures. The tool made it possible to create reference columns with

original values, apply GREL expressions, perform batch edits, identify duplicates, and cluster terms using different methods. Clustering was employed as a decision-support mechanism, rather than as an automatic replacement operation. Each grouping suggestion was assessed by the team, taking into account the meaning of the term, the source metadata, the museological context, the collection, the object class, and the risk of information loss.

Curatorial validation was a central stage of the process. The decisions involved professionals with knowledge of documentation processes, the collections, cataloging practices, and the internal uses of the database. This validation made it possible to distinguish syntactic problems, such as unnecessary spaces and capitalization variations, from semantic problems, such as the mixture of information in inappropriate fields, ambiguous use of terms, and the need to create or redefine metadata.

Finally, the process was documented in reports and decision records, with the aim of preserving the institutional memory of the choices made, guiding future data imports, and supporting the update of the cataloging manual.

4.0 Results

The analysis of the metadata from the Three-Dimensional and Iconographic collections showed that the main problems were not limited to isolated spelling or formatting errors, but were associated with the historical accumulation of documentary practices carried out over approximately three decades. Among the problems identified, the most significant were terminological dispersion, multi-valued entries in inappropriate fields, the mixture of semantically distinct information within the same metadata element, the use of non-standardized abbreviations, data types incompatible with the informational function of the field, and values that required redistribution to more appropriate metadata elements.

The first relevant result was the construction of a systematic diagnosis of the database. The classification of metadata according to low, medium, and high levels of effort made it possible to establish intervention priorities and to distinguish automatable treatments from those requiring specialized curatorial analysis. This classification was not based solely on the number of terms existing in each field, but on the relationship among semantic dispersion, consistency of field completion, and structural complexity. Thus, the methodology adopted demonstrated that metadata quality depends on both syntactic and semantic aspects.

Among the metadata elements treated, ‘Other periods,’ from the Three-Dimensional collection, constitutes a significant case of controlled reduction in terminological dispersion. The treatment reduced the number of distinct terms from 928 to 657, corresponding to a 29.2% reduction in the vocabulary used in the field. This reduction did not result from indiscriminate automatic substitutions, but from an analysis of the meaning of the recorded values, distinguishing historical periods, names of rulers, minting dates, commemorative information, and contextual observations. Values that did not correspond to the informational function of the metadata element were migrated to more appropriate

fields, such as “Date notes” or “History.” Thus, the quantitative reduction was accompanied by semantic reorganization.

Another central result was the restructuring of location metadata. The analysis revealed that information related to the current storage location, temporary movements, previous records, and the history of object transfers was distributed unclearly across different fields. Based on validation with the museum team, a distinction was consolidated among “Permanent Location,” understood as the current physical storage location; “Temporary Location,” intended for situations of transit, exhibition, or loan; and “Location History,” aimed at preserving the chronological memory of movements. This reorganization shows that metadata quality depends on the functional clarity of fields: when a metadata element accumulates distinct functions, its consistency and retrievability are impaired.

In agent metadata, such as “Manufacturer,” “Signatory,” and “Designer/Engraver,” the results demonstrated the need to balance terminological normalization with the preservation of context. In “Manufacturer,” names accompanied by locations, addresses, expressions of uncertainty, abbreviations, and institutional information were identified. The team chose not to automatically eliminate these elements, since some of them could hold historical or documentary relevance. In “Signatory,” by contrast, the treatment led to a significant reduction in terminological dispersion, from approximately 1,408 to 762 distinct terms. The process involved the standardization of personal and institutional names, as well as the treatment of honorific titles, positions, legal representatives, estates, and acquisition information. In “Designer/Engraver,” the treatment required attention to the different authorial functions recorded in the field, such as designer, engraver, photographer, obverse, and reverse.

The use of OpenRefine was decisive in enabling treatment at scale. The tool made it possible to create reference columns, apply GREL expressions, perform textual transformations, identify duplicates, execute batch edits, and generate grouping suggestions. However, clustering was used as a decision-support instrument, not as an automatic replacement mechanism. Each suggestion was analyzed by the team, taking into account the museological context, the function of the metadata element, the collection of origin, and the risk of information loss. This procedure confirms the need to combine automatic metrics and procedures with specialized human judgment.

In light of the data quality literature, the results show that qualifying the database cannot be understood merely as a search for accuracy. As Wang and Strong (1996) argue, high-quality data are data that are fit for use. In the case of the Museu Paulista, many values were not necessarily incorrect, but were located in inappropriate fields, hindering interpretation, retrieval, or reuse. An old location, for example, could be historically true but inappropriate as the current permanent location; a name accompanied by a position could be informative, but unsuitable as the main value of an agent metadata element. Thus, the treatment sought to align the data with the informational function of each field.

From the perspective of metadata quality, the results engage with dimensions such as consistency, coherence, completeness, accessibility, and conformity to expectations, as discussed by Bruce and Hillmann (2004) and Park (2009). Consistency was expanded

through the reduction of terminological variants and the standardization of values. Coherence was strengthened by the redistribution of information to semantically appropriate fields. Accessibility and retrievability were favored by the elimination of textual noise, the normalization of vocabularies, and the improvement of record intelligibility. In addition, the creation of reference columns and the systematic documentation of decisions contributed to the traceability of the process and to the preservation of institutional memory.

These results allow metadata curation to be understood as an applied practice of Knowledge Organization. According to Hjørland (2008), Knowledge Organization involves both processes and systems of representation. In the case analyzed, the treatment included processes of cleaning, normalization, validation, and reorganization, but also affected systems of representation, such as taxonomies, selection lists, controlled vocabularies, and the institutional thesaurus. The main contribution of the process, therefore, lies not only in the numerical reduction of terms, although this reduction is relevant, but in the consolidation of an institutional methodology for information curation capable of articulating computational tools, quality criteria, museological knowledge, and documentation of decisions.

5.0 Conclusion

The study has limitations. The first concerns the specificity of the case analyzed, since the Museu Paulista has its own institutional trajectory, legacy documentary database, cataloging practices, and technical team, which prevents the automatic generalization of the results to other cultural heritage repositories. In addition, the metadata curation process is still ongoing, meaning that not all metadata elements have been fully treated or validated. Another limitation refers to the absence, at this stage, of direct evaluation with end users, which would make it possible to measure more precisely the impacts of the treatment on the experience of searching, browsing, and retrieving information.

Despite these limitations, the study contributes by demonstrating that metadata curation in museological repositories should not be understood merely as the technical cleaning of data, but as an applied practice of Knowledge Organization. The analysis showed that diagnostic assessment, level-of-effort classification, terminological normalization, clustering, curatorial validation, and the conceptual reorganization of fields can expand the semantic consistency and quality of metadata. From a methodological standpoint, the main contribution lies in the systematization of a replicable workflow, although one that necessarily depends on adaptation to each context.

Enhancing the quality of the Museu Paulista metadata constitutes a fundamental step toward the sustainability of its digital repository, the improvement of information retrieval, and future interoperability initiatives. The experience demonstrates that the migration of legacy databases to digital platforms must be accompanied by continuous data review, treatment, and governance. Metadata curation thus emerges as a strategic activity for cultural heritage institutions, since it makes it possible to reconcile the preservation of

documentary memory, informational consistency, public access, and the production of knowledge about collections.

AI Usage Statement: AI was used in this paper to assist with text suggestions and corrections. In the preparation of the English version of this manuscript, generative AI tools were used as auxiliary support for checking lexical choices, terminological consistency, and academic phrasing by the translator.

References

Ballou, Donald P., Richard Y. Wang, Harold Pazer, and Giri Kumar Tayi. 1998. "Modeling Information Manufacturing Systems to Determine Information Product Quality." Management Science 44, no. 4: 462–484. https://doi.org/10.1287/mnsc.44.4.462.

Bruce, Thomas R., and Diane I. Hillmann. 2004. "The Continuum of Metadata Quality: Defining, Expressing, Exploiting." In Metadata in Practice, edited by Diane I. Hillmann and Elaine L. Westbrooks, 238–256. Chicago: American Library Association.

Gilliland, Anne J. 2016. "Setting the Stage." In Introduction to Metadata, 3rd ed., edited by Murtha Baca. Los Angeles: Getty Publications. https://www.getty.edu/publications/intrometadata/setting-the-stage/.

Guimarães, José Augusto Chaves. 2008. "A dimensão teórica do tratamento temático da informação e suas interlocuções com o universo científico da International Society for Knowledge Organization (ISKO)." Revista Ibero-Americana de Ciência da Informação 1, no. 1: 77–99. https://doi.org/10.26512/rici.v1.n1.2008.940.

Hjørland, Birger. 2008. "What Is Knowledge Organization (KO)?" Knowledge Organization 35, nos. 2–3: 86–101. https://doi.org/10.5771/0943-7444-2008-2-3-86.

Hjørland, Birger. 2017. "Domain Analysis." Knowledge Organization 44, no. 6: 436–464. https://doi.org/10.5771/0943-7444-2017-6-436.

Ingwersen, Peter, and Kalervo Järvelin. 2005. The Turn: Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer.

Mey, Eliane Serrão Alves. 1995. Introdução à catalogação. Brasília: Briquet de Lemos.

Novellino, Maria Salet Ferreira. 1996. "Instrumentos e metodologias de representação da informação." Informação & Informação 1, no. 2: 37–45. https://doi.org/10.5433/1981-8920.1996v1n2p37.

Ochoa, Xavier, and Erik Duval. 2009. "Automatic Evaluation of Metadata Quality in Digital Repositories." International Journal on Digital Libraries 10, nos. 2–3: 67–91. https://doi.org/10.1007/s00799-009-0054-4.

Park, Jung-ran. 2009. “Metadata Quality in Digital Repositories: A Survey of the Current State of the Art.” Cataloging & Classification Quarterly 47, nos. 3–4: 213–228. https://doi.org/10.1080/01639370902737240.

Pipino, Leo L., Yang W. Lee, and Richard Y. Wang. 2002. “Data Quality Assessment.” Communications of the ACM 45, no. 4: 211–218. https://doi.org/10.1145/505248.506010.

Saracevic, Tefko. 1999. “Information Science.” Journal of the American Society for Information Science 50, no. 12: 1051–1063. https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1051::AID-ASI2>3.0.CO;2-Z.

Wang, Richard Y., and Diane M. Strong. 1996. “Beyond Accuracy: What Data Quality Means to Data Consumers.” Journal of Management Information Systems 12, no. 4: 5–33. https://doi.org/10.1080/07421222.1996.11518099.