Bridging the Gap Between Curated and Induced Semantics
C. Shimizu et al. (Eds.)

doi:10.3233/SSW260020

From File Systems to Knowledge Graphs: A Multi-Phase Workflow for the Description of Born-Digital Literary Archives

Lucia GIAGNOLINI [b,1] Paolo BONORA [a,b]

[a] *Alma Mater Studiorum – Università di Bologna, Dipartimento di Filologia Classica e Italianistica (FICLIT), Bologna, Italy*

[b] *Digital Humanities Advanced Research Center (/DH.arc), Alma Mater Studiorum – Università di Bologna, Bologna, Italy*

ORCiD ID: Lucia Giagnolini https://orcid.org/0000-0002-4876-2691, Paolo Bonora https://orcid.org/0000-0001-8337-3379

Abstract. Background: Born-digital personal archives present novel descriptive challenges that conventional archival methods are not equipped to address. Their scale, format heterogeneity, and multi-layered stratification — physical, logical, and conceptual — require a representational framework capable of capturing document structure, provenance, integrity, and relational contexts. The increasing volume of born-digital archives received by cultural heritage institutions makes the development of automated and scalable description methodologies an urgent practical priority. Methods: This paper presents a replicable five-phase pipeline, grounded in the Born-Digital Ontology (BoDi) — a formal extension of Records in Contexts Ontology (RiC-O) integrating PREMIS, PROV-O, and LRMoo — that transforms born-digital personal archive directories into RDF knowledge bases. The phases are: (1) file system ingestion, comprising SHA-256 integrity verification, parallel multi-tool metadata extraction, and triplestore loading; (2) SPARQL-based validation and rule-based enrichment, including cross-tool metadata alignment, document classification, and duplicate detection; (3) hybrid enrichment combining expert bibliographic curation with automated LLM-based description generation; (4) chain-of-custody reconstruction; and (5) selective entity-level redaction for privacy-aware dissemination. Results: Evaluated on the Valerio Evangelisti Archive — 78,211 files and 11,135 folders totaling 2.1 TB across three storage media — the pipeline generated 61,154,322 RDF triples across 13 named graphs with minimal manual intervention. The resulting knowledge base supports querying across structural, typological, temporal, and relational dimensions, enabling analyses — from duplicate detection across storage media to correlation of archival activity with editorial output — that conventional finding aids cannot support. Conclusions: Born-digital archives present a double challenge: their fragility raises preservation concerns, while their scale makes manual description almost impossible. The pipeline addresses both by inverting the framing: the properties that make born-digital materials resistant to conventional methods become, within a formal framework, sources

[1] Corresponding Author: Lucia Giagnolini, lucia.giagnolini2@unibo.it

of descriptive richness without equivalent in the analog domain. Ontological modeling combined with automated processing and human-in-the-loop validation offers a viable and empirically demonstrated path toward semantic description of born-digital cultural heritage at scale.

Keywords. born-digital archives, knowledge graphs, RDF, semantic enrichment, provenance, retrieval-augmented generation, archival description, BoDi ontology, Records in Contexts

1. Introduction

Contemporary authorship generates complex, distributed, and stratified documentary ecosystems. Resulting digital materials — manuscripts at successive revision stages, email correspondence, research notes, image libraries, executable scripts, and compressed archives — accumulate across multiple digital storage media over decades, often without deliberate organizational principles and in the absence of any preservation plan [1]. This is particularly true of literary archives, where the documentary ecosystem reflects decades of layered creative activity whose documentary significance is inseparable from its temporal and relational complexity [2]. Born-digital personal archives therefore reach collecting institutions in a state of minimal pre-arrangement, placing the full descriptive burden on archivists and digital preservation specialists.

Traditional archival approach frames the digital as an obstacle, interpreting the technical complexity of the medium for a descriptive problem — a perspective that holds only when born-digital archives are viewed through the lens of analog methodologies. In reality, born-digital archives are computational objects whose digital nature enable automated description with a level of granularity that paper records cannot support. Unlike analog documents, digital records carry intrinsic descriptive layers — format signatures, system metadata, hierarchical file system topologies — that can be accessed and extracted directly [3]. The hierarchical structure of the file system, moreover, encodes organizational and topological relationships that can be captured with full fidelity — something that analog finding aids can only approximate through reconstruction. Born-digital materials carry extensive machine-readable traces of their production and organization. The archival challenge is therefore not to generate descriptive metadata *ex nihilo*, but to extract, formalize, and semantically enrich the information already available in the digital record.

Translating this reframing into practice requires both a robust representational model and a scalable computational framework. Current standards adopted by Libraries, Archives and Museums communities [4,5,6,7,8] still leave a significant gap between accurate technical information and archival description; and a single author's digital heritage may span tens of thousands of files, making manual entity-level description unfeasible at scale.

This paper addresses the following research question: *how can the intrinsic properties of digital materials be leveraged to automate and facilitate archival description processes?* To the best of our knowledge, existing approaches address this question only partially, lacking the integration of a robust representational model and a scalable, reproducible pipeline within a single framework. We address this question with three contributions: (1) a five-phase, openly available pipeline that transforms born-digital archive di-

rectories into RDF knowledge graphs with minimal manual intervention, grounded in the Born-Digital Ontology (BoDi)[2]; (2) a hybrid enrichment methodology combining rule-based inference, expert curation, and local LLM-based generation; and (3) a large-scale empirical validation on a real-world archive. The pipeline covers structure and metadata extraction and validation, semantic enrichment, automated description generation, chain-of-custody reconstruction, and privacy-preserving redaction.

The remainder of this paper is organized as follows. Section 2 reviews representational requirements, existing models, and the Born-Digital Ontology as the framework that underpins the pipeline. Section 3 describes the phases of the pipeline. Section 4 reports the experimental evaluation on the Valerio Evangelisti Archive. Section 5 discusses findings and limitations. Section 6 concludes and outlines the directions for future work.

2. Background and Related Work

Representing born-digital personal archives requires addressing a set of challenges that arise from the specific nature of digital objects and contemporary documentary production [9]. Born-digital materials exhibit an inherently stratified nature: following Thibodeau [10], every digital object manifests three distinct, yet interconnected, levels of abstraction — physical (bit sequences on storage hardware), logical (file system structure), and conceptual (informational content) — each of which must be explicitly represented. Beyond stratification, digital integrity cannot be assumed and must be cyclically verified: the cryptographic hash of a file provides an objective, machine-verifiable record of its bit-level identity at a given point in time [11]. At the same time, born-digital files accumulate heterogeneous technical metadata automatically during their life-cycle, generated by systems, applications, and users across different technological environments [12,13] — a layer of information whose variability and granularity cannot be anticipated in advance, requiring a flexible, tool-agnostic representational approach. Finally, the provenance of digital materials is multidimensional, encompassing technical, procedural, and curatorial dimensions [14]: the full chain of actors, tools, and operations that produced, transformed, and managed the archive must be documented, as must the relational contexts — bibliographic, social, intellectual — that situate documents within the broader ecosystem of authorial production.

These requirements find a natural match in Linked Data, which the international archival community has progressively adopted as its representational paradigm [15,4] for its flexibility, relationality, and extensibility. The Records in Contexts conceptual model (RiC-CM) and its OWL ontology (RiC-O) [4] represent the most advanced vocabulary for archival linked data, consolidating ISAD(G) and ISAAR(CPF) within a graph-based framework capable of representing complex networks of archival entities, agents, and contexts. RiC-O is progressively affirming itself as the new international reference standard for archival description. Alongside it, domain-specific ontologies such as ArCo [8] and ArchOnto [16] address particular institutional contexts, while the digital preservation community relies on PREMIS [5] for technical metadata and fixity at the object level.

Nevertheless, none of these models satisfies the above requirements simultaneously. As a general-purpose archival ontology, RiC-O does not natively implement the techni-

[2] http://w3id.org/bodi

cal layers required for born-digital description — a gap that EGAD itself acknowledges, noting that RiC-O "is designed with the expectation that it may be extended using cross-domain ontologies specifically addressing technical data needed for the management of instantiations" [4]. PREMIS addresses fixity at the preservation object level, but lacks the conceptual archival description layers needed to situate digital objects within archival, biographical, and intellectual contexts. ArCo is scoped to the Italian cultural heritage context and does not address born-digital peculiarities properly; CIDOC CRM, while internationally adopted, is designed for museum-oriented object description and similarly leaves some of the born-digital technical characterization unaddressed.

The same gap persists at the operational level. Tools such as Archivematica[3], BitCurator[4], and format identification utilities such as DROID[5] solve the problems of archive ingest, fixity verification, format normalization, and OAIS-compliant packaging. Their output is serialized partly in PREMIS, but its scope is deliberately bounded to preservation-relevant properties, leaving the broader technical characterization of digital objects outside its representational range. More fundamentally, these systems are oriented toward conservation and access rather than description. They identify a digital object and whether it has been altered, but they do not define the contents, how it relates to other objects, or where it sits within the archival, biographical, and intellectual context of its creator.

Closer to the domain of born-digital literary archives, the Pavia Archivi Digitali (PAD) system [1] implements automated metadata extraction from multiple tools and provides browsing functionality, but stores its data in a relational database without modeling according to archival ontologies or standard vocabularies, limiting its capacity for querying and interoperability. The present work proceeds from the preservation layer into the descriptive domain: from identifying the digital object, to the semantic characterization of its contents, to the relation with other objects, and finally where it sits within the archival, biographical, and intellectual context of its creator — a representational depth that preservation systems are not designed to reach.

The Born-Digital Ontology (BoDi) was developed within the broader research program underlying this work to fill this gap; it is taken here as a premise, established prior to and independently of the present contribution. Taking RiC-O as its reference framework, BoDi selectively integrates PREMIS for fixity and storage medium representation, PROV-O [17] for the provenance of processing operations, and LRMoo [6] for the bibliographic layer connecting archival records to literary works. Where these vocabularies do not cover the representational requirements of born-digital description, BoDi introduces new classes and properties grounded in three design principles. First, extracted metadata values are represented as explicit entities in the knowledge graph, while metadata fields themselves (e.g., the File Size) are modeled as entities rather than just scalar attributes.

This design enables provenance tracking, semantic clustering, and cross-tool alignment at the level of metadata concepts while preserving the original terminology used by each extraction tool. Equivalent metadata fields extracted by different tools can therefore be semantically aligned to their original extraction context — a granularity that neither RiC-O nor PREMIS supports natively. The additional graph complexity is a delib-

[3] https://www.archivematica.org/en/

[4] https://bitcurator.net/

[5] https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

erate trade-off that enables provenance-aware comparison of heterogeneous extraction outputs.

Second, the provenance model is extended to cover computational agents — algorithms, software components, and composite software environments — alongside human agents, with explicit properties for documenting supervisory relationships in hybrid human-machine workflows; unlike RiC-O, whose provenance properties are scoped to archival entities, BoDi's corresponding properties apply to any domain entity, enabling provenance tracking of computational outputs such as extracted metadata values and automatically generated descriptions.

As part of the broader research programme in which BoDi was developed, the ontology has been formally validated through RDF syntax checking, OWL 2 DL consistency reasoning (HermiT), and structural quality assessment (OOPS! [18]), confirming logical coherence at both the terminological and assertional levels [9].

The design principle is modularity: RiC-O remains fully usable as a standalone framework, and the imported ontologies are employed only where they directly address representational needs that RiC-O alone does not cover.

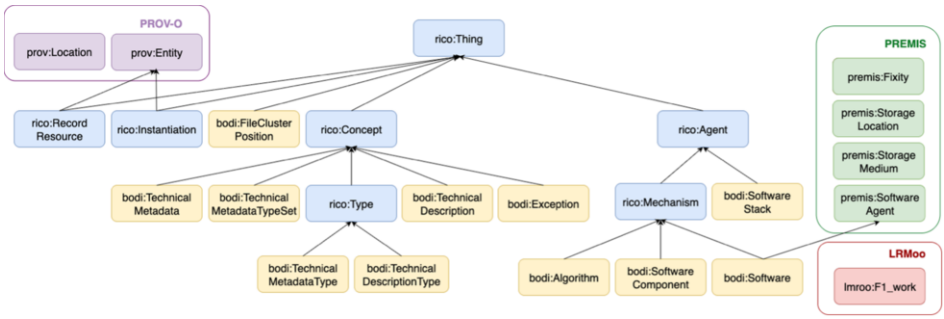

Figure 1. The Born-Digital Ontology (BoDi): modular integration of RiC-O, PREMIS, PROV-O, and LRMoo, with BoDi-native classes and properties highlighted.

3. The Five-Phase Pipeline

The pipeline takes as input one or more directories of a born-digital archive and produces as output a multi-graph RDF knowledge base conformant with BoDi. It is implemented in Python 3.12, orchestrated by a central script, and uses Blazegraph as the target triplestore, accessed via REST APIs. All data is serialised in N-Quads format, ensuring portability across any W3C-compliant triplestore.[6]

The pipeline is designed around three methodological commitments. First, (1) *integrity preservation*: every operation maintains the physical and logical integrity of the source materials through cryptographic verification and preventive backups. Second, (2) *reversibility and transparency*: all transformations are documented and traceable, connecting the files to their semantic representations through verifiable, reproducible processes. Third, (3) *standard compliance*: all generated representations conform to docu-

[6]The code is publicly available at https://github.com/LuciaGiagnolini12/FileSystem_to_BoDi.

mented models — BoDi, RiC-O, PREMIS, PROV-O, LRMoo — guaranteeing portability, interoperability, and reusability. Table 1 summarizes the five phases.

Table 1. Overview of the five-phase pipeline.

Phase	Key Process	Output
1. Archive Ingestion and RDF Baseline Construction	Read-only protection; SHA-256 integrity baseline; RDF conversion of file system hierarchy; parallel metadata extraction via three tools; SPARQL consistency and integrity verification	N structural + $3N$ metadata named graphs; named graph registry; JSON hash registers
2. Validation & Rule-Based Enrichment	SPARQL validation suite (entity counts, hierarchical consistency, hash integrity, metadata coverage); rule-based enrichment: duplicate detection, cross-tool alignment, metadata clustering, type assignment, title and date formalization	Validated KG; named graph of duplicate links, cross-tool equivalences, type assignments, formalized titles and dates
3. Semantic Enrichment via Specialized Knowledge and Generative Models	Expert bibliographic mapping to literary work entities; deterministic SPARQL retrieval feeding a locally executed LLM for natural-language description generation	Work-entity named graph; named graph of natural-language descriptions with provenance
4. Chain of Custody Reconstruction	Physical media census; storage medium and location modeling; derivation relations between instantiations with temporal and agent anchoring	Named graph of storage media, locations, and derivation sequences
5. Selective Redaction for Publication	Configurable whitelist/blacklist policy by entity type; field-level screening of sensitive metadata; redaction status flag on all entities	Redacted KG with all entities flagged; full pre-redaction backup

3.1. Phase 1: Archive Ingestion and RDF Baseline Construction

Phase 1 transforms one or more file system directories into an RDF knowledge base. It is the only phase requiring direct access to the file system; all subsequent phases operate exclusively within the knowledge base. The phase proceeds through ten sequential steps, each completing across all target directories before the next begins; a dependency system blocks advancement unless all prerequisite artifacts are present, ensuring full traceability.

Read-only protection. All archive directories are set to `read-only` mode before any operation begins. The system automatically generates a directory-specific recovery script for each target, guaranteeing complete reversibility.

Integrity baseline. A SHA-256 cryptographic fingerprint (hash) is computed for every file using native OS commands (`sha256sum` on Linux). Results are serialized in structured JSON registers recording hash values, file sizes, timestamps, and processing statistics. This baseline serves three purposes: (1) attesting the state of the source mate-

rials at the moment of ingestion; (2) detecting any subsequent modification of those materials throughout the workflow; and (3) enabling identification of duplicate files across directories (see Phase 2).

RDF structure generation. The file system hierarchy is converted into a BoDi-conformant RDF representation. Files and folders are modeled as `rico:Record` and `rico:RecordSet` instances, respectively, each paired with a `rico:Instantiation` representing their filesystem realization and each situated in its hierarchical context according to its position in the file system. Hash values are recorded as `premis:Fixity` entities, file paths as `prov:Location` instances. The resulting triples are stored in a dedicated named graph per directory and loaded into the triplestore via the REST API.

Quality control. Two independent SPARQL-based checks are executed before proceeding to metadata extraction. A *consistency check* verifies that entity counts in the RDF graph match those obtained from Bash file system commands; an *integrity check* recomputes the SHA-256 values and verify the matching with those in the JSON registers of the Integrity baseline.

Multi-tool metadata extraction. Native technical metadata is extracted through three specialized tools. The Python `os` library[7] extracts file system metadata (timestamps, permissions, inode identifiers). Apache Tika[8] provides format-agnostic content metadata across over 1000 file formats, mapping extracted fields onto standard vocabularies including Dublin Core and XMP. ExifTool[9] targets multimedia files, extracting deep technical parameters such as EXIF data, codec specifications, audio frequencies, and GPS coordinates. Each tool's output is stored in a dedicated named graph per directory, so that the provenance of every metadata value is always traceable to its source extractor. Running the three tools in parallel maximizes coverage: any surface properties the others may not capture, or cases where the same property is extracted by multiple tools with differing values, are preserved for cross-validation [3].

Post-extraction integrity verification. Once metadata extraction is complete, SHA-256 values of all source files are newly recomputed and compared against the Integrity baseline. Any discrepancy is flagged and logged, certifying that no source file was altered during processing.

Named graph registry. A final named graph formalizes the dependencies between all named graphs produced across directories and extraction tools. It functions as a centralized index that facilitates identification of graphs relevant to specific queries and supports targeted maintenance operations, preserving modularity while providing a navigable view of the full knowledge base.

3.2. Phase 2: Validation and Rule-Based Enrichment

Phase 2 operates exclusively within the semantic domain on the knowledge base produced by Phase 1. It comprises two sequential sub-processes: systematic validation of the graph's logical consistency, followed by rule-based enrichment that makes implicit knowledge explicit. All triples generated in this phase are stored in a dedicated named graph (`updated_relations`), keeping them separate from the Phase 1 baseline and ensuring complete reversibility.

[7] https://docs.python.org/3/library/os.html
[8] https://tika.apache.org/
[9] https://exif.tools/

Validation is conducted through a suite of 51 SPARQL queries, formulated as competency questions covering entity distributions and label coverage, structural and hierarchical integrity, metadata extraction coverage, and hash compliance and duplication. The set was derived directly from the entities and relations defined by the BoDi/RiC-O data model, each query targeting a specific structural or referential invariant that Phase 1 output is expected to satisfy; it is not claimed to be formally complete, but designed to cover the main risk categories in born-digital ingestion identified in Section 2.

Rule-based enrichment builds on the outputs of the informational queries to generate new triples through six operations that surface knowledge already latent in the data:

(1) *Hash-based duplicate detection*: files sharing identical SHA-256 values across input directories are linked to each other surfacing duplication patterns and potential backup trajectories across media.
(2) *Cross-tool metadata alignment*: metadata type labels produced by the three extraction tools (`os` library, Apache Tika, and ExifTool) were reviewed by a digital archivist with expertise in digital preservation, who identified 96 semantically equivalent pairs by triangulating lexical similarity of field names, cross-tool agreement of extracted values on the same files, and consistency with each tool's technical documentation. The pairs of metadata types are formalized as `owl:sameAs` relations and loaded into the knowledge graph.
(3) *Semantic clustering*: metadata types are grouped into nine functional categories covering file system attributes, document content, image, audio, video, email, executable, archive, and security metadata, broadening the navigational granularity of the knowledge base.
(4) *Document type assignment*: Major media types (MIME) are mapped to human-readable labels and assigned to each file's archival entity, improving the readability and filterability of the collection.
(5) *Title formalization*: file and folder names are modeled as dedicated title entities according to the RiC-O model, enabling richer title metadata.
(6) *Date formalization*: temporal metadata is normalized and modeled as dedicated date entities, distinguishing dates pertaining to the intellectual content of a record from dates pertaining to its physical digital carrier, and encoded in both machine-readable and natural-language representations.

3.3. Phase 3: Semantic Enrichment via Specialized Knowledge and Generative Models

Phase 3 extends the knowledge base beyond what can be inferred from the archive itself, introducing knowledge from two external sources: a specialist bibliographic expertise and a locally executed generative language model.

Literary work mapping. A domain expert conducts a bibliographic survey of the author's works, identifying individual titles and their eventual groupings into cycles, series, or trilogies — categories illustrative of literary production and not an exhaustive or fixed taxonomy: for archives of a different nature, this stage can adapted by substituting the relevant grouping unit while retaining the same part-of modeling pattern. These are modeled as work entities following the LRMoo ontology [6], with part-of relations encoding the hierarchy between cycles and their constituent works. Individual archival records and record sets are then linked to the relevant work entity; hierarchical propa-

gation ensures that a record associated with a specific work is also implicitly associated with any cycle or series containing it. This integration enables content-aware querying across the knowledge base. All generated triples are stored in a dedicated named graph. In the Evangelisti case study, the bibliographic survey and mapping of 30 works into 8 cycles and trilogies was carried out by a single domain expert over approximately three working days, a cost that scales with the size and complexity of the author's bibliography rather than with archive size.

Automatic description generation. For each file in the archive, the system automatically generates a concise natural-language description by combining deterministic retrieval from the knowledge graph with a locally executed LLM. For each archival instantiation, a SPARQL query retrieves all associated technical metadata, organized by semantic priority: identifying properties first, followed by structural and descriptive fields. This structured context is passed to Llama 3.2 3B [19], executed locally via Ollama, with a prompt instructing the model to write a brief, factual, 2–3 sentence description based only on the supplied metadata, without speculation, and with units and dates normalized (the full prompt template is available in the code repository).

The system shares some characteristics with retrieval-augmented generation approaches, while relying on deterministic graph retrieval rather than vector similarity. It differs from conventional KGQA systems by producing narrative output from structured input, rather than structured tuples from narrative queries. Given the sensitive nature of these archival materials, local execution is a requirement rather than a preference: no file content or metadata is transmitted to external infrastructures, and the use of open-source model weights ensures full auditability and reproducibility.

An illustrative example from the case study shows the system's interpretation capability. Given the following technical metadata for an RTF document:

```
Content-Type: application/rtf  |  Pages: 5  |  Words: 1739
CreateDate: 2008:02:15 12:00:00  |  ModifyDate: 2008:07:29 04:43:00
RevisionNumber: 33  |  TotalEditTime: 6.0 hours
LastModifiedBy: Valerio Evangelisti  |  st_size: 100735
```

The system generates: *"Rich Text Format (RTF) document, 5 pages with 1,739 words. Created by Valerio Evangelisti in February 2008, revised through 33 versions with 6 hours total editing time, last modified by Valerio Evangelisti in July 2008. Titled '3', likely a chapter or section of a larger work. File size 98.4 KB."*

The output demonstrates unit conversion (bytes to KB), date normalization, and contextual inference from structured metadata. Generation parameters (temperature $= 0.3$, `num_predict` $= 150$, `top_p` $= 0.9$) favor descriptive consistency over lexical variety. Each generated description is annotated with full provenance — generating activity, model identity, supervising agent, and timestamp — and flagged for potential human review. Descriptions are inserted incrementally with checkpoint persistence, enabling resumption after interruption, and stored in a dedicated named graph.

Both processes are independently configurable: the scope of the work mapping and the subset of records targeted for description generation can be adjusted to the specific archive and access policy. In this way, the two processes can also be made interdependent — as in the Evangelisti case study, where description generation is restricted to the records made visible by the work mapping step, focusing computational resources on the materials that will be effectively accessible to users.

3.4. Phase 4: Chain of Custody Reconstruction

Phase 4 reconstructs and formalizes the physical chain of custody connecting the original storage media to the working copies used for processing. Documenting this history is archivistically essential: each transfer, copy, or migration operation may have left traces in the files themselves — through system-specific metadata, format conversions, or file system timestamps — and the awareness of who handled the materials, under what conditions, and through which technological environments is necessary to interpret anomalies and assess the reliability of the archive as received. In the absence of prior inventories — a frequent condition in personal archives — the phase begins with a manual census of all identified storage supports, recording medium type, provenance, current location, and conservation conditions. This census provides the empirical grounding for all subsequent modeling.

Each storage medium and its physical location are represented as dedicated entities using the PREMIS vocabulary, extended by BoDi to connect them to the corresponding archival instantiations in the knowledge graph. Derivation relationships between successive instantiations — documenting copy, migration, or transfer operations from one medium to another — are modeled explicitly, with each relationship temporally anchored to the date of the operation and linked to the agents involved.

The resulting named graph transforms the custodial history from background knowledge held by archivists into queryable, machine-readable provenance — making it possible to connect every file in the archive to the full sequence of hands, media, and environments it passed through before reaching the researcher.

3.5. Phase 5: Selective Redaction for Publication

Phase 5 prepares the knowledge base for dissemination by replacing sensitive label strings with a standardized placeholder rather than removing entities, preserving the graph's relational structure while concealing sensitive content; a full backup precedes any operation, ensuring reversibility. Sensitive data in born-digital personal archives is distributed not only in file contents but also in metadata fields — creator names, last-author values, file paths — requiring field-level, not just record-level, screening.

Redaction follows differentiated, configurable logic per entity type. Archival records are redacted by default and remain visible only if they satisfy one of three criteria agreed with rights holders (in the Evangelisti case, with the Associazione Valerio Evangelisti – Il Sol dell'Avvenire): explicit linkage to a literary work (`rico:isRelatedTo` to an `lrmoo:F1_Work`), inclusion in an authorization list, or hierarchical containment (`rico:isOrWasIncludedIn`) in an authorized `rico:RecordSet`. Matching is implemented as a SPARQL classification against these properties; entities queued for redaction have their label and title replaced via SPARQL `DELETE/INSERT` with the placeholder *"Redacted information"*, and every entity is marked with a boolean `bodi:redactedInformation` flag, providing a queryable audit trail. Folders follow the inverse logic (visible unless blacklisted). For visible entities, metadata fields referring to third parties are additionally screened and redacted field-by-field. Defining these criteria with the archive's custodian was an iterative, one-time policy decision independent of archive size. Post-redaction consistency checks verify that there is no mismatch between entities and their instantiations.

4. Results

4.1. Dataset and Infrastructure

The pipeline was applied to the born-digital partition of the Valerio Evangelisti (1952–2022) Archive — an Italian science-fiction and historical-fiction writer — including materials from three storage media: a copy of the main Windows PC hard disk, an external hard disk (Western Digital My Book 2TB), and the extracted contents of 93 floppy disks, spanning approximately 1984–2022 and totaling 2.1 TB.

Phase 1, in particular, was conducted on a Dell PowerEdge R7525 server (2× AMD EPYC 7313 16-core CPUs, 256 GB RAM) running Ubuntu 24.04.3 LTS, using Python 3.12.3 with `rdflib` 7.1.4, `SPARQLWrapper` 2.0.0, and Apache Tika 3.2.1. The pipeline generated 61,154,322 RDF triples across 13 named graphs, stored in a Blazegraph journal of 18.33 GB. Blazegraph was deployed with default configuration settings and without query optimization; the reported performance figures should therefore be interpreted as a reproducible baseline rather than an optimized upper bound. Table 2 shows the distribution of archival entities across storage media.

Table 2. Distribution of archival entities across storage media.

Storage medium	Files	Folders
Main hard disk	56,171	9,624
External hard disk	19,810	1,394
Floppy disks	2,230	117
Total	**78,211**	**11,135**

4.2. Knowledge Graph Quality

The knowledge graph represents 78,211 files and 11,135 folders, each carrying one fixity and one location entity. Technical metadata extraction yielded 5,574,418 metadata entities, distributed across the three tools as follows: Apache Tika 2,221,401 values, ExifTool 2,191,519 values, `os` library 1,161,498 values. The tools exhibit markedly different and complementary type coverage. The `os` library produces a stable set of 13 metadata types per directory, capturing folder-level attributes unavailable to the other tools. Apache Tika yields 960 distinct types on the main hard disk, 867 on the external hard disk, and 244 on the floppy corpus, reflecting progressively narrower format diversity. ExifTool surfaces up to 106,211 distinct types on the main hard disk, reflecting the exceptional density of multimedia content. No single extractor would have captured the full descriptive depth represented in the knowledge base alone. Post-extraction integrity verification confirmed that all SHA-256 values were identical before and after extraction, providing evidence that the source materials remained unaltered during processing.

The 51-query SPARQL validation suite was executed against the complete knowledge base, completing in 403 seconds with 100% execution success (0 errors, 0 failures). The validation suite detected no structural violations across any functional group: no hierarchical inconsistencies, no orphan entities, no metadata values without provenance, no hash format violations, no multiple hash values per file, no invalid file paths. 349 files

remain classified as `application/unknown` or `application/octet-stream` — a known limitation of MIME-based classification for corrupted or legacy binary formats, consistent with the format obsolescence documented in the floppy disks corpus. Query execution times range from under 5 ms for simple entity counts to 44 seconds for full provenance coverage scans over metadata entities, demonstrating acceptable SPARQL performance on a dataset of this scale without query optimization.

4.3. Semantic Enrichment

Expert-curated cross-tool alignment produced 96 equivalence pairs covering properties from file size and timestamps to camera model, audio sample rate, PDF version, font family, and GPS coordinates, enabling unified cross-tool querying. Hash-based duplicate detection identified 9,080 hash groups shared across files, involving 35,577 file instances in total; the most replicated hash (1,816 occurrences) corresponds to the universal empty-file value, surfacing a systematic pattern of placeholder or migration artifacts that would be entirely invisible in a traditional finding aid. Structural analysis revealed a maximum nesting depth of 18 folder levels on the main hard disk, with the densest concentration of files between levels 9 and 12. MIME type classification identified 60 distinct types across 57,015 classified files; images and documents together account for over 86% of the corpus, with RTF as the dominant document format (9,319 files), reflecting a sustained use of legacy word-processing environments. The most frequent types are reported in Table 3.

Table 3. Most frequent media types in the Valerio Evangelisti Archive.

MIME Type	Occurrences
`image/jpeg`	17,778
`text/rtf`	9,319
`text/plain`	6,204
`image/png`	4,933
`image/svg+xml`	3,462
Other (55+ distinct types)	15,319

Literary work mapping linked 6,073 archival entities to 30 works organized into 8 narrative cycles and trilogies, enabling content-aware querying across the knowledge base. The combination of work linkage and temporal metadata formalization made it possible to correlate archival activity with the author's editorial output. For example, the biennium 2005–2006 emerges as the peak of computational activity in the archive, coinciding with the publication of two novels and a short story collection — a correlation that the knowledge graph surfaces automatically from metadata alone, without access to file contents.

The full transfer history of each storage medium from its original physical location to the working copies used for processing. Three derivation sequences are recorded — one for the main hard disk, one for the external hard disk, and one for the floppy disk corpus. The resulting named graph allows to trace, for any file in the archive, the sequence of hands, media, and environments it passed through before reaching the processing infrastructure.

4.4. Description Generation Quality

The automatic description system produced 5,548 natural-language technical descriptions in 3 hours and 46 minutes on a consumer laptop (Apple M1 Pro, 16 GB RAM) running Llama 3.2 3B locally via Ollama. As a preliminary quality assessment, a domain expert reviewed a sample of 100 descriptions covering the five most frequent MIME types, evaluating each against its source metadata on three criteria: factual correctness of all stated properties, absence of hallucinated content, and appropriate unit conversion and date normalization. All 100 descriptions passed all three criteria. This evaluation confirms the basic factual adequacy of the generation system; a more systematic validation protocol — addressing sample size, single-evaluator design, and binary criteria — is planned as outlined in Section 5. Manual production of equivalent descriptions is estimated at 5–10 minutes per file, placing the equivalent manual effort between 58 and 116 person-working-days.

4.5. Knowledge Base Deployment

The resulting knowledge base is deployed on a ResearchSpace [20] instance dedicated to the Valerio Evangelisti Archive, providing a browser-based interface for SPARQL querying and visual exploration of the archive, built on top of the knowledge base produced by the pipeline presented here and described separately in [21]. Navigation is organized around a file system metaphor, with every visualization generated by live SPARQL queries executed directly against the underlying knowledge graph. Given the sensitive nature of the materials, access is currently restricted pending final review by the Associazione Valerio Evangelisti – Il Sol dell'Avvenire, custodian of the archive and of the rights of its contents.

5. Discussion

5.1. A Shift in Archival Epistemology

The pipeline and BoDi together instantiate a shift in how archival description of born-digital materials can be approached: rather than requiring an archivist to inspect and document each item individually, the pipeline extracts, formalizes, and connects the descriptive information already embedded in the digital objects, producing a queryable knowledge graph that provides a rich, structured baseline for targeted curation and expert intervention.

This changes the archival workflow rather than merely accelerating it: producing technical descriptions of thousands of files individually is operationally impossible, so the archivist is freed to focus on interpretation and contextualization — bibliographic references, access policy, editorial judgment — while the bottleneck shifts from producing descriptions to verifying them, and from comprehensive coverage to targeted intervention.

Scale itself becomes an interpretive resource rather than merely a technical challenge: the analyses reported in Section 4 — from cross-media duplicate detection to the correlation of archival activity with editorial output — are the kind of insight that scalable reading [22] affords, moving fluidly between macro-level patterns and targeted close

reading, with the knowledge graph functioning as a navigation instrument rather than a static inventory.

This granularity is a structural consequence of the KG format, not merely a byproduct of automation. Preservation-oriented formats (e.g., OAIS packages, PREMIS-only records) safeguard bit-level integrity but remain opaque to cross-entity querying; relational approaches such as PAD (Section 2) would need a proliferation of join tables to capture the same cross-cutting, many-to-many relations that the graph model represents natively. No manually curated finding aid reaches a comparable level of detail — a scale of curation that contemporary born-digital archives render impracticable. A formal, comparative evaluation of this added value against a traditional finding aid has not yet been conducted and is identified as future work below.

5.2. Limitations and Future Work

The most structural limitation affects Phase 5. The current architecture encodes access policies through static whitelists and blacklists, effectively reproducing the binary logic of paper-based access control within a digital environment. While operationally sufficient, this approach does not fully exploit the expressiveness of the knowledge graph. A graph-native access architecture based on differentiated named-graph layers and role-based visibility policies would provide flexible and maintainable solution, allowing access conditions to be updated independently from the processing pipeline itself. More importantly, such an approach would fully exploit the relational nature of the graph, allowing researchers with different authorization levels to navigate the same knowledge base at different degrees of resolution [23].

The pipeline has been validated on the Valerio Evangelisti Archive, to our knowledge the largest known born-digital personal archive preserved by an Italian cultural heritage institution; although validation relies mainly on a single case study, the archive's exceptional heterogeneity in formats, typologies, and organizational structure provides a demanding testbed. Phases 1, 2, and 4 are fully archive-agnostic, while Phases 3 and 5 require adaptation to the target collection, while the underlying modeling pattern and the rest of the pipeline remain structurally independent from the Evangelisti case. Computationally, Phases 1 and 2 scale linearly with archive size, while Phase 3 (description generation) and SPARQL query execution — ranging from 5 ms to 44 s on the current dataset without optimization — are the principal bottlenecks for substantially larger corpora, motivating planned parallelization, graph partitioning, and query caching. Broader empirical validation is ongoing: beyond tests on smaller sample archives, the pipeline is currently being applied, at a preliminary stage, to a second, structurally different archive — the unallocated but migrated files of the Gabriel García Márquez archive — expected to provide further evidence of the pipeline's behavior outside the Evangelisti Archive conditions.

Additional limitations concern metadata enrichment and semantic classification. The literary work mapping currently depends on domain expertise and prior bibliographic reconstruction, limiting transferability to archives lacking comparable scholarly resources. MIME type classification also remains intrinsically limited by the semantic ambiguity of certain media types; content-level disambiguation combining OCR and layout analysis is therefore envisaged as a complementary layer of interpretation.

The knowledge base is not yet publicly accessible, pending final review by the Associazione Valerio Evangelisti – Il Sol dell'Avvenire, the custodian of the archive.

The preliminary evaluation of automatically generated descriptions discussed in Section 4 likewise requires substantial extension. The planned protocol will assess a stratified sample of 300–500 descriptions across MIME types and metadata-density levels, independently reviewed by an archivist (factual accuracy, disciplinary adequacy) and a non-specialist user (accessibility, clarity). Evaluation will move from a binary pass/fail model to a Likert-scale framework covering factual accuracy, completeness, and linguistic quality, complemented by automatic checks of formally verifiable properties (date formats, unit conversions).

6. Conclusion

The five-phase pipeline presented in this paper shows the practical viability of large-scale semantic description for born-digital personal archives. Applied to the Valerio Evangelisti Archive, the system transformed a corpus whose manual processing would have required years of sustained archival work into a machine-queryable knowledge environment.

The significance of the contribution, however, lies not only in automation or scale, but in the underlying shift discussed in Section 5: from document-centric description toward a relational, computationally explorable model. At the same time, the work highlights the extent to which born-digital archival processing remains an unresolved methodological and infrastructural problem. Questions of access control, semantic classification, provenance modeling, evaluation methodology, and computational scalability require further experimentation across archives of different provenance and institutional context. The present study should therefore be understood not as a definitive solution, but as a scalable operational framework and a proof of methodological feasibility for knowledge-graph-based archival description.

More broadly, born-digital archives are not simply digital equivalents of paper fonds requiring new preservation strategies, but fundamentally different documentary environments whose complexity can only be adequately represented through relational and machine-processable models.

Author Contributions

Lucia Giagnolini designed and developed the pipeline, conducted the case study, and wrote the manuscript. Paolo Bonora supervised the work and contributed to the design of the pipeline.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to perform language editing and restructuring. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Giagnolini L., Baldini P. Literary Bytes: New Approaches to Born-Digital Archives at Pavia Archivi Digitali. In: Antonelli G., Giagnolini L., Milone F., editors. The Intangible Papers. Authorial Philology and Born-Digital Texts. Bologna: Il Mulino; 2025. p. 57–78. doi: 10.1401/9788815414328/c3.

[2] Carbé E. Digitale d'autore. Macchine, archivi, letterature. Firenze: Firenze University Press – USiena Press; 2023.

[3] Gorini A., Giagnolini L. Under the Surface: Interpreting Technical Metadata as Archival Narrative. In: Proceedings of the 21st International Conference on Digital Preservation (iPRES 2025); 2025; Wellington, New Zealand. https://phaidra.univie.ac.at/o:2198090

[4] ICA Experts Group on Archival Description. Records in Contexts Ontology (RiC-O), version 1.0. International Council on Archives; 2023. https://www.ica.org/standards/RiC/ontology

[5] Caplan P. Understanding PREMIS. Library of Congress Network Development and MARC Standards Office; 2009. https://www.loc.gov/standards/premis/understanding-premis.pdf

[6] IFLA LRMOO Working Group. IFLA Library Reference Model Object-Oriented (LRMoo): Definition and Application, version 0.9.6. International Federation of Library Associations; 2024. https://www.iflastandards.info/lrmoo

[7] Bekiari C., Bruseker G., Doerr M., Ore C.E., Stead S., Velios A., editors. Definition of the CIDOC Conceptual Reference Model, version 7.2. CIDOC CRM Special Interest Group; 2021. https://www.cidoc-crm.org/

[8] Carriero V.A., Gangemi A., Mancinelli M.L., et al. ArCo: The Italian Cultural Heritage Knowledge Graph. In: Ghidini C., et al., editors. Proceedings of the 18th International Semantic Web Conference (ISWC 2019); 2019 Oct 26–30; Auckland, New Zealand. Cham: Springer; 2019. p. 36–52. doi: 10.1007/978-3-030-30796-7_3.

[9] Giagnolini L. Rappresentare il digitale d'autore: dal file system al knowledge graph. PhD thesis, Alma Mater Studiorum Università di Bologna. Dottorato di ricerca in Patrimonio culturale nell'ecosistema digitale, 38 Ciclo; 2026. doi: 10.48676/unibo/amsdottorato/13078.

[10] Thibodeau K. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In: The State of Digital Preservation: An International Perspective. Washington (DC): Council on Library and Information Resources; 2002. p. 4–31.

[11] Digital Preservation Coalition. Fixity and Checksums. In: Digital Preservation Handbook, 2nd ed.; 2015. https://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums

[12] Larry D., Lars D. Digital Forensics for Legal Professionals: Understanding Digital Evidence From the Warrant to the Courtroom. Burlington: Syngress; 2011. https://www.ebsco.com/ebooks

[13] The Sedona Conference. Commentary on Ethics & Metadata. Sedona Conference Journal. 2013;14:169.

[14] Bak G. Digital Provenance. Archival Science. 2024;24(4):847–69. doi: 10.1007/s10502-024-09462-w.

[15] Hyvönen E. Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Morgan & Claypool Publishers; 2012. doi: 10.2200/S00452ED1V01Y201210WBE003.

[16] Koch I., Ribeiro C., Teixeira Lopes C. ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives. In: Hall M., Merčun T., Risse T., Duchateau F., editors. Digital Libraries for Open Knowledge. Lecture Notes in Computer Science, vol. 12246. Cham: Springer International Publishing; 2020. doi: 10.1007/978-3-030-54956-5_10.

[17] Lebo T., Sahoo S., McGuinness D., et al. PROV-O: The PROV Ontology. W3C Recommendation; 2013. https://www.w3.org/TR/prov-o/

[18] Poveda-Villalón M., Gómez-Pérez A., Suárez-Figueroa M.C. OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. International Journal on Semantic Web and Information Systems (IJSWIS). 2014;10(2):7–34.

[19] Dubey A., Jauhri A., Pandey A., et al. The Llama 3 Herd of Models. arXiv:2407.21783; 2024.

[20] Oldman D., Tanase D. Reshaping the Knowledge Graph by Connecting Researchers, Insights and Practices in ResearchSpace. In: Vrandečić D., et al., editors. Proceedings of the 17th International Semantic Web Conference (ISWC 2018); 2018 Oct 8–12; Monterey, CA. Cham: Springer; 2018. p. 325–340. doi: 10.1007/978-3-030-00668-6_20.

[21] Grillo R., Giagnolini L., Bonora P. Archival Knowledge Graphs: Balancing Semantic Richness and Accessibility of Hierarchical Structures. In: Proceedings of the 22nd Italian Research Conference on Digital Libraries (IRCDL 2026); 2026 Feb 19–20; Modena, Italy. CEUR Workshop Proceedings; 2026. In press.

[22] Krautter B. The Scales of (Computational) Literary Studies: Martin Mueller's Concept of Scalable Reading in Theory and Practice. In: Armaselu F., Fickers A., editors. Exploring Scale in Digital History and Humanities. Berlin: De Gruyter Oldenbourg; 2024. doi: 10.1515/9783111317779-011.

[23] Kirrane S., Mileo A., Decker S. Access control and the resource description framework: A survey. *Semantic Web*. 2017;8(2):311–352. doi: 10.3233/SW-160236.