

A Multi-Purpose Model for Structural Manuscript Data Storage and Representation

Marta Riparante¹[0009-0006-5701-9804], Maxim Goynov²[0009-0007-2244-4314],
Detelin Luchev²[0000-0003-0926-5796], Desislava Paneva-Marinova²[0000-0001-5998-687X],
Lilia Pavlova²[0009-0006-2777-9386]

¹ Cyrillo-Methodian Research Centre, Bulgarian Academy of Sciences,
25 Acad. Georgi Bonchev Street, 1113 Sofia, Bulgaria

² Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
8 Acad. Georgi Bonchev Street, 1113 Sofia, Bulgaria
martariparante.bas@gmail.com, m.goynov@math.bas.bg,
dml@math.bas.bg, dessi@cc.bas.bg,
pavlova.lilia@gmail.com

Abstract. This paper presents a hybrid data-model architecture developed for the Repertorium of South Slavonic Manuscripts and Copyists, a digital system for the description and analysis of Cyrillic manuscripts produced in the South Slavic lands. The model functions simultaneously as a structured storage schema, an ontological model for knowledge management, and a framework governing data representation and interface behaviour. By integrating codicological description and palaeographical analysis within a single machine-readable definition, it reduces schema–interface divergence, lowers maintenance overhead, and improves data consistency. The case study shows how domain-driven modelling can support rigorous and operationally efficient digital infrastructures for manuscript studies.

Keywords: Manuscript Studies, Data Storage, South Slavic Cyrillic Manuscripts, Digital Humanities.

1 Introduction

One of the central challenges in digital humanities infrastructure is designing systems that can accommodate complex, domain-specific scholarly data while remaining accessible and practical for researchers engaged in analysis and publication (van Lit, 2020; Verhoeven & Burrows, 2015). The conventional architectural response to this challenge is the separation of data models from presentation layers, considered as a principle valued for promoting modularity and long-term maintainability (Sanderson, 2024). However, when applied to large, dynamic data structures subject to continuous scholarly revision, strict adherence to this separation introduces coordination overhead and heightens the risk of divergence between layers, with potential consequences for data

integrity (Martin-Rodilla & Gonzalez-Perez, 2019). Despite this paradigmatic separation, the interface and the underlying data model remain deeply intertwined in practice, as observed by Herbst et al. (2023).

This separation manifests in practice as two parallel artifacts: a schema defining entities, fields, and relationships, and a separate presentation layer that maps these structures to input forms, display templates, and search interfaces. Any modification to the descriptive framework – adding a new field, reclassifying a typology, or refining a controlled vocabulary – requires synchronized updates across both layers. Empirical studies of digital humanities infrastructure have documented how this dual maintenance burden can lead to gradual schema–presentation drift, where the interface no longer faithfully represents the underlying data model, or conversely, where the data model accumulates legacy fields that are never exposed to researchers. In codicological and palaeographic scholarship, where manuscript description categories are themselves objects of ongoing methodological debate (Amory, 2025; Horobin, 2024; Kestemont et al., 2017; Korogodina & Levchenko, 2025; Orsini, 2025), this risk is particularly acute.

This paper examines an alternative approach implemented in the *Repertorium of South Slavonic Manuscripts and Copyists (10th–14th c.)* (hereafter referred to as Repertorium), (<https://kopisti.kmnc.bg/en>), a digital humanities system developed within research project “Fourteenth Century South Slavonic Scribes and Scriptoria (Palaeographical Attribution and Online Repertorium)” (No KP-06-N50/4/30.11.2020) and further developed within “Development of the Cyrillic alphabet from the 9th to the 14th century in the South Slavic lands: research and digital presentation” (No KP-06-N90/9/11.12.2024), dedicated to the description, storage, and intelligent curation of medieval South Slavonic manuscripts (Luchev et al., 2024; Scarpa et al., 2021). The system serves palaeographers, codicologists, and philologists working on manuscripts preserved in repositories worldwide. Rather than delegating presentation and interaction logic to a separate layer, the model simultaneously defines a structured framework for manuscript description and encodes the display behaviour, interface configuration, and multilingual rendering required by researchers – a consolidation that proves particularly advantageous where the cost of inter-layer divergence exceeds the cost of architectural coupling.

2 The Repertorium

The Repertorium is a digital humanities project dedicated to the systematic description and scholarly analysis of Cyrillic manuscripts produced in the South Slavic lands. The system serves as both a research database and a publication platform, enabling scholars to record, query, and disseminate structured descriptions of manuscript witnesses, scribal hands, and textual transmission within this geographic and historical scope.

The current version of the data model at the core of the system reflects the complexity of this domain and has been shaped in response to ongoing scholarly and technical developments. A single ‘Manuscript’ record is organized across three top-level descriptive domains, encompassing dozens of nested descriptive categories. The ‘Identification’ domain captures the primary referential and classificatory metadata through which

each manuscript unit is located and recognized within the system. It includes the hierarchical place of preservation (country, region, repository, with standardized and original forms of institutional names, abbreviated library sigla, and website link), specific collection and shelfmark data, the indication of codicological part with folio range where applicable, as well as the manuscript's nickname and its typological classification according to textual content. The 'Description' domain covers the manuscript's material and codicological characteristics, containing information on date and provenance, material, the number of folia and endpapers, folio dimensions, written area and text field measurements, number of columns, writing lines, and the relevant bibliography, together with accompanying notes where necessary. The 'Palaeography' domain records script-level features. It includes information on orthographic tradition, script type, copyist attributions, and the identification and description of individual hands, together with colophons and other scribal notes where present, and records handwriting aspect, letterforms, and other graphic elements, as well as links to manuscript images and image-based documentation. Each domain is itself composed of structured sub-objects with typed fields, controlled vocabularies, and cross-references to independent entities such as 'Copyists', 'Locations', and 'Sources'. This three-level organization reflects the methodological structure of manuscript scholarship directly: the categories are not generic metadata fields but theoretically determined observational units whose precise articulation affects the analytical conclusions that can be drawn from the data.

3 The Repertorium Model

Given this descriptive complexity, a conventional two-layer architecture, in which a data schema and a presentation layer are defined and maintained independently, would introduce synchronization overhead disproportionate to the scale of the development team and the frequency of model revisions. The Repertorium addresses this through a hybrid design in which a single definition serves both purposes simultaneously.

At the structural level, the model (Fig. 1) defines a multilevel hierarchy, spanning from manuscript to description, layout, and individual field measurements, in which each node carries typed properties (text, number, radio, switch), validation constraints, and relationship references to external entities such as copyists, locations, and sources. Key fields participate in both data retrieval and display logic, while computed fields (e.g., "compute":"cacheCentury") enable derived attributes that support search and presentation without additional processing layers.

Crucially, the same definition also encodes presentation behaviour. Field labels are declared bilingually inline (e.g., "name":"Име", "name:en":"Name"), eliminating the need for external localization files. Properties such as "presFn":"skipEmptyObject" govern rendering logic, while "cClass":"flex-h no-collapse" controls layout directly through CSS class assignments. Contextual hints for data-entry editors are embedded at the field level, keeping instructional content co-located with the data it describes.


```

        "id": "part",
        "name": "Часть",
        "type": "text",
        "name:en": "Part",
        "key": "scalar"
    }, {
        "id": "from-folio",
        "name": "От лист",
        "type": "text",
        "name:en": "From Folio",
        "key": null
    }, {
        "id": "to-folio",
        "name": "До лист",
        "name:en": "To Folio",
        "type": "text"
    }, {
        "id": "notes",
        "name": "Бележки",
        "type": "textarea:ml",
        "name:en": "Notes"
    }
],
"cls": true,
"presFn": "skipEmptyObject",
"cClass": "flex-h no-collapse"
},

```

This single object simultaneously specifies what data is collected for the description of manuscript parts, in what form it is entered, which field functions as its scalar identifier, and how the object is labelled and rendered in the interface in different languages. The same structure therefore encodes both the scholarly category itself and the practical rules governing its display and use. The result is a self-describing structure in which schema and presentation are not merely compatible but constitutively linked. While this coupling trades some architectural flexibility for operational coherence, it represents a deliberate and well-motivated design choice for systems where scholarly precision and continuous model evolution demand that structure and representation remain perpetually in sync.

4 Methodological Implications

The architecture of the Repertorium model carries broader methodological implications that extend beyond its specific implementation. Most directly, it exemplifies domain-

driven modelling: the structure of the data is not determined by generic software conventions but by the analytical requirements of manuscript scholarship itself. The underlying schema-and-presentation mechanism is uniform across domains, while the disciplinary categorisation it encodes differs in maturity. The codicological component implements the descriptive framework established for the Repertorium in Paneva-Marinova et al. (2022), whereas the palaeographical component implements the analytical model proposed in Riparante (2024), whose categories continue to be refined as the methodology for the palaeographic analysis of South Slavonic Cyrillic develops. Because structure and presentation are encoded within the same definition, such refinements can be incorporated incrementally, without requiring parallel updates to a separate presentation layer – a concrete instance of the maintenance advantage discussed in Section 1. Each metadata element corresponds to a discrete analytical point of interest, producing a model that is simultaneously theoretically grounded and operationally useful for researchers.

A second implication concerns the elimination of translation layers. Conventional architectures require explicit mapping between data schemas and user-interface definitions, a process that introduces both development overhead and a persistent risk of inconsistency whenever either layer is modified. By encoding structural and presentational logic within the same definition, the hybrid model removes this translation requirement entirely. Researchers interact directly with a representation of their own conceptual framework, rather than with an interface that approximates it through a separate rendering layer.

The integration also has measurable consequences for data quality. Contextual hints embedded at the field level guide editors toward correct input formats at the point of entry. Conditional display logic, including role-based field visibility and custom presentation functions, reduces cognitive load and minimizes input errors. Type-specific validation (numeric constraints, controlled vocabularies, boolean and date datatypes) enforces structural integrity without requiring post hoc correction. Taken together, these mechanisms ensure that data is collected in a format that is both flexible enough to accommodate scholarly nuance and strict enough to support systematic analysis.

5 Comparative Analysis

Measured against conventional layered architectures, the hybrid model presents a distinct trade-off profile. The key differences are summarized in Table 1.

Table 1. Traditional Approach vs. Hybrid Model

Feature	Traditional Approach	Hybrid Model
Model complexity	High (two independent models)	Moderate (single integrated model)
Data consistency	Requires cross-layer validation	Intrinsic to single definition

Development efficiency	Lower (parallel development tracks)	Higher (single model to maintain)
Domain fidelity	Often compromised by abstraction	High (domain drives structure)
Maintenance overhead	Higher (two models to synchronize)	Lower (single point of change)
Team scalability	Suited to large, specialized teams	Viable for small teams or individuals

The pattern that emerges is not simply that the hybrid model is superior, but that it optimizes for a different set of priorities. Traditional separation maximizes modularity and scalability across large development teams, at the cost of coordination complexity and potential layer divergence. The hybrid model accepts tighter coupling in exchange for operational coherence and domain fidelity – a trade-off that is well-suited to specialized scholarly systems where the data model is inherently unstable and the development team is small.

6 Conclusion

The Repertorium model demonstrates that in specialized scholarly domains, the conventional separation between data structures and presentation layers is not a neutral architectural choice but an assumption that must be evaluated against the nature of the research object itself. In manuscript studies, this issue is particularly pressing. Medieval manuscripts are not merely carriers of textual content, but historically produced and culturally stratified artefacts whose analysis requires the coordinated perspectives of codicology, palaeography, philology, and textual scholarship. A single codex may simultaneously preserve evidence of material construction, textual transmission, scribal intervention, later additions, and broader cultural practices of use and preservation. Any digital model intended to represent such objects must therefore remain closely aligned with the categories through which scholars actually observe and interpret them.

From this perspective, the hybrid approach examined here addresses a concrete methodological and technical problem. Where descriptive frameworks are revised in response to ongoing research, maintaining two synchronized layers – one for data structure and one for presentation – can generate disproportionate overhead and introduce avoidable risks to data integrity. The Repertorium responds by treating the data model as a self-describing artefact, one that encodes not only what data exists, but also how it should be entered, validated, displayed, and interpreted within a discipline-specific framework. In this respect, the model is especially well suited to manuscript studies, because it accommodates both the codicological description of the codex as a physical and textual unit and the palaeographical analysis of script, hands, and graphic features, while preserving their methodological distinction within a common structure.

The ongoing refinement of the palaeographical component, discussed in Section 4, is not incidental to this argument but constitutes a live test of the architecture’s resili-

ence under active disciplinary revision. The categories used for palaeographic description are currently being updated in line with the methodology under development for the *Cyrillic Palaeography Toolkit* (CyPaT), an online platform for the description, structuring, processing, management, visualisation, and comparative analysis of codicological and palaeographical data derived from digitised images of Cyrillic manuscripts (Riparante et al., 2025). While the physical characteristics of the folio remain part of the ‘Manuscript’ object within the Repertorium, CyPaT supports the detailed analysis of script features relevant to the ‘Palaeography’ domain, including measurements of letter-height-to-nib-width ratios, writing and text angles, and stroke analysis based on a standardised palaeographic vocabulary, alongside more advanced functions such as vector annotation, pattern matching, and letter classification. As these analytical categories evolve through CyPaT’s development, the hybrid model must accommodate them without disrupting the relationship between structure and presentation – a demonstration, in practice, of the maintenance advantage outlined above, and an indication that the case examined here is not an isolated exception but representative of how the architecture is intended to behave under continuous scholarly revision.

More broadly, this case suggests that the design of digital humanities infrastructures should begin not from abstract architectural orthodoxy, but from the epistemic demands of the material under study. For projects dealing with complex manuscript artefacts, a unified model may offer a more faithful, transparent, and maintainable representation of scholarly knowledge than a rigidly layered architecture. Rather than asking by default how data and presentation should be separated, developers of domain-specific infrastructures might more productively ask at what point such separation ceases to serve the object of research – and whether a model shaped directly by the logic of manuscript scholarship provides a more adequate foundation for long-term scholarly work.

Acknowledgements

This research work was carried out and is funded partly by the Bulgarian Scientific Fund under the research project No KP-06-N90/9/11.12.2024 “Development of the Cyrillic alphabet from the 9th to the 14th century in the South Slavic lands: research and digital presentation”.

References

- Amory, Y. (2025). More than a simple intuition: towards a categorization of palaeographical features in Greek documentary papyri. *Comparative Oriental Manuscript Studies Bulletin*, 1–18. <http://hdl.handle.net/1854/LU-01HXXGDHQ4JPCP5D0MKEC4HS57>
- Herbst, Y. W., Roeder, T., & Sadari, S. M. S. (2023). *On the Relationship Between Interfaces and TEI in Digital Scholarly Editions*. Text Encoding Initiative. <https://old.tei-c.org/Vault/MembersMeetings/2023/contributions/149.html>
- Horobin, S. (2024). Identifying Scribal Hands: Principles and Problems. *Speculum*, 99(3), 688–696. <https://dx.doi.org/10.1086/730563>

- Kestemont, M., Christlein, V., & Stutzmann, D. (2017). Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts. *Speculum*, 92(S1), S86–S109. <https://doi.org/10.1086/694112>
- Korogodina, M. V., & Levchenko, N. N. (2025). Konvoliut ili sbornik: problemy tipologizatsii [Convolute or miscellany: the problem of typology]. *Peterburgskii istoricheskii zhurnal*, 1(45), 155–164. https://doi.org/10.51255/2311-603X_2025_1_155
- Luchev, D., Goynov, M., Paneva-Marinova, D., Pavlov, R., & Rangochev, K. (2024). Repertoire of medieval South Slavic manuscripts and scribes in research context. *Chuzhdoezikovo Obuchenie – Foreign Language Teaching*, 51(1), 75–89. <https://doi.org/10.53656/for2024-01-09>
- Martin-Rodilla, P., & Gonzalez-Perez, C. (2019). Metainformation scenarios in Digital Humanities: Characterization and conceptual modelling strategies. *Information Systems*, 84, 29–48. <https://doi.org/10.1016/j.is.2019.04.009>
- Orsini, P. (2025). The Dating of Early Christian Manuscripts: Paleography and Material Culture. In S. W. Crawford & T. Wasserman (Eds.), *The Oxford Handbook of Textual Criticism of the Bible* (pp. 419–445). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780197581315.013.0025>
- Paneva-Marinova, D., Scarpa, M., Riparante, M., Goynov, M., Luchev, D., Pavlova, L., & Zlatkov, L. (2022). Machine-readable descriptive structures to study medieval South Slavic scriptoria and scribes. *TEM Journal*, 11(2), 675–682. <https://doi.org/10.18421/TEM112-22>
- Riparante, M. (2024). Predlozhenie za paleografski analiz na srednovekovnata yuzhnoslavyanskata kirilitsa [A Proposal for the Palaeographic Analysis of Medieval South Slavonic Cyrillic Script]. *Palaeobulgarica*, 48(2), 113–127. <https://doi.org/10.59076/2603-2899.2024.2.06>
- Riparante, M., Goynov, M., Paneva-Marinova, D., & Pavlova, L. (2025). Enhancing South Slavic Cyrillic Manuscripts Research through a Digital Toolkit for Cyrillic Palaeography. In S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, & F. Remondino (Eds.), *Digital Heritage*. The Eurographics Association. <https://doi.org/10.2312/dh.20253058>
- Sanderson, R. (2024). Implementing Linked Art in a Multi-Modal Database for Cross-Collection Discovery. *Open Library of Humanities*, 10(2), 1–25. <https://doi.org/10.16995/olh.15407>
- Scarpa, M., Burlacu, C., & Paneva-Marinova, D. (2021). New digital life for fourteenth century South Slavonic scriptoria. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, 11, 57–64. <https://doi.org/10.55630/dipp.2021.11.5>
- van Lit, L. W. C. (2020). *Among digitized manuscripts: Philology, codicology, paleography in a digital world*. Brill.
- Verhoeven, D., & Burrows, T. (2015). Aggregating cultural heritage data for research use: The Humanities Networked Infrastructure (HuNI). In *Metadata and Semantics Research: 9th Research Conference, MTSR 2015, Proceedings* (pp. 445–452). Springer. <https://doi.org/10.1007/978-3-319-24129-6>

Received: March 15, 2026

Reviewed: May 10, 2026
Finally Accepted: May 30, 2026