A semantic interoperability model based on NLP for non-structured health data

Blanda Helena De Mello[1], **Sandro José Rigo**[1], **Cristiano André da Costa**[1]

[1]Laboratório de Inovação em desenvolvimento de Software – SOFTWARELAB
Programa de Computação Aplicada–Universidade UNISINOS
Av. Unisinos 950 – São Leopoldo – 93022-750 – Brasil.

{rigo, cac}@unisinos.br, blanda@edu.unisinos.br

Abstract. *The increasing volume of unstructured clinical data challenges interoperability, especially in decentralized systems like SUS. This study proposes a semantic model integrating NLP, machine learning, and ontologies to extract and standardize clinical notes. Using Named Entity Recognition (NER), lexical normalization, and alignment with HL7 FHIR and openEHR, the model was tested on COVID-19 data from partner hospitals, demonstrating its effectiveness in structuring unstructured data and enabling scalable interoperability.*

1. INTRODUCTION

The digital transformation of healthcare is a global priority, enhancing system efficiency and accessibility. Interoperability standards, such as HL7 FHIR and ICD-11, play a crucial role in integrating health information systems, ensuring seamless data exchange and continuity of care [da Silva Jr et al. 2024]. However, the COVID-19 pandemic exposed critical gaps in data management, highlighting the urgency of structuring and standardizing clinical information across different healthcare levels [Sim et al. 2023]. As healthcare institutions increasingly adopt Electronic Health Records (EHR), fragmented governance, infrastructure disparities, and the lack of semantic standardization hinder interoperability [Mougin et al. 2022]. The 2024-2030 PAHO Plan for Strengthening Health Information Systems (IS4H) emphasizes emerging technologies and AI-driven solutions to address these challenges, reinforcing the need for harmonized health data models.

Despite digital health advancements, clinical data remains highly fragmented, collected across multiple care levels, services, and specialties, complicating data exchange between public and private sectors [Sivarethinamohan et al. 2021]. This heterogeneity limits secondary applications such as epidemiological surveillance, clinical research, and decision support. Achieving sustainable digital transformation requires structured governance, investment in interoperability, and workforce capacity building. The IS4H plan further promotes cross-border data integration through the Pan American Highway for Digital Health (PH4H), aligning with the need for semantic models capable of structuring unstructured clinical narratives.

In this context, this research proposes a model for semantic interoperability, integrating Natural Language Processing (NLP), ontologies, and machine learning techniques to extract, structure, and standardize clinical data. The model leverages structured vocabularies and controlled terminologies, ensuring data harmonization across healthcare institutions. By transforming unstructured clinical records into a machine-readable format, it facilitates information exchange, interoperability, and scalable integration, supporting key applications in clinical research, public health monitoring, and healthcare system optimization.

2. THEORETICAL BACKGROUND

Interoperability ensures data consistency and seamless integration across health systems [Sheth 1999]. The Healthcare Information and Management Systems Society (HIMSS) defines four levels—Foundational, Structural, Semantic, and Organizational[HIMSS 2021]—to address exchange challenges. The Semantic level is particularly crucial in healthcare, enabling interpretable data exchange through standardized terminologies and ontologies[Benson and Grieve 2021]. Despite the growing adoption of Electronic Health Records (EHRs), data fragmentation and unstructured textual fields remain obstacles to interoperability[ISO18308 2011]. The transition from paper-based to digital records has introduced usability concerns, highlighting the need for automated semantic structuring to optimize clinical documentation and data exchange[Martin-Sanchez and Verspoor 2014].

Semantic interoperability ensures that health data retains its meaning across systems[Alexopoulos 2020]. However, healthcare data is inherently heterogeneous, collected across institutions and specialties, creating integration challenges. The absence of structured vocabularies and normalization techniques leads to inconsistencies, limiting data reuse for clinical decision-making and public health[El Kah and Zeroual 2021]. The Organizational level further emphasizes governance and policy compliance in digital health[Benson and Grieve 2021].

The exponential growth of unstructured clinical data has driven the adoption of Natural Language Processing (NLP) and Machine Learning (ML) in healthcare[Shen et al. 2021]. NLP extracts and structures medical text, supporting semantic interoperability[Martin-Sanchez and Verspoor 2014]. However, challenges remain due to medical terminology complexity, clinical documentation variability, and a lack of annotated corpora[Hasan and Farri 2019]. Advances in Deep Learning (DL), particularly Transformer-based models, have improved semantic structuring, fostering interoperability[Li et al. 2021]. Named Entity Recognition (NER) plays a key role in structuring clinical data, extracting entities such as diseases, drugs, and procedures[Li et al. 2020]. However, fragmented data from clinical notes, imaging reports, and prescriptions require semantic harmonization[Martin-Sanchez and Verspoor 2014]. The integration of ML and NLP techniques enhances entity normalization, supporting data-driven decision-making, interoperability, and scalable digital health solutions[Li et al. 2021, Li et al. 2020].

3. DESIGN AND IMPLEMENTATION OF THE MODEL

Healthcare data is often fragmented, stored in diverse formats, and recorded as unstructured clinical narratives, posing interoperability challenges. This study introduces a semantic interoperability model that processes unstructured healthcare data, integrating natural language processing (NLP), machine learning, and ontologies to standardize clinical information and facilitate structured data exchange. The model is developed under the MyDigitalHealth project, in collaboration with six hospitals in Brazil. The approach ensures interoperability by aligning data with HL7 FHIR, openEHR, SNOMED CT, and

ICD-10. The methodological pipeline consists of five stages: (1) data acquisition, (2) preprocessing and anonymization, (3) annotation and corpus creation, (4) named entity recognition (NER) and information extraction, and (5) ontology-based representation. Data includes clinical notes from hospitalized COVID-19 patients, collected under the Research Ethics Committee protocol CAAE: 33540520.6.3004.5327.

A fine-tuned BERTimbau model was applied to anonymize personally identifiable information (PII) before annotation, ensuring privacy compliance. The annotation process involved medical students and specialists, using the UBIAI tool to label six entity categories: Clinical Attributes, Disease or Syndrome, Signs or Symptoms, Diagnostic Procedure, Invasive or Therapeutic Procedure, and Test Result. For entity extraction, we evaluated three Transformer-based models: BERT-multilingual-cased, BERT-Portuguese-cased, and BioBERTpt-Clin. Models were assessed using Precision, Recall, and F1-score, with a 70%-15%-15% data split. Lexical normalization techniques were applied to improve entity recognition and consistency. The extracted entities were structured into an OWL ontology, aligned with CIDO[1], CODO[2], and Covid-O[3], ensuring semantic harmonization. The ontology development process followed a structured approach: requirement specification, modeling, validation, and evaluation. This representation bridges unstructured clinical narratives and interoperability standards, enabling structured data integration. By leveraging NLP and ontology-based modeling, the proposed pipeline supports automated knowledge extraction, semantic interoperability, and scalable clinical data exchange. The following section details the proposed model's architecture and its role in advancing healthcare interoperability.

This research[4] presents a semantic interoperability model developed under the MyDigitalHealth project, designed to address interoperability gaps in electronic health records (EHRs) where clinical narratives and structured data coexist. The model was developed in collaboration with hospital partners, ensuring alignment with real-world healthcare data challenges. The model follows a layered approach based on [HIMSS 2021], covering: (i) Foundational layer: Addresses technical infrastructure and electronic system integration; (ii) Structural layer: Ensures compatibility by defining a healthcare standard to interoperate, such as HL7 FHIR, openEHR, and terminologies SNOMED CT, ICD, and others; (iii) Semantic layer: Uses Named Entity Recognition (NER) and ontologies to standardize extracted terms, matching terms on existing terminologies; (iv) Organizational layer: Governs data-sharing policies and compliance.

To process unstructured clinical narratives, the model integrates a structured pipeline for information extraction. Using Transformer-based models like BERT, it extracts and normalizes key medical terms, ensuring alignment with standardized medical terminologies. Lexical normalization mitigates variability in free-text entries, improving data consistency. A key aspect of the model is its modular design, allowing flexible adoption by institutions with varying levels of digital maturity. The extracted entities are structured in an OWL ontology, serving as an intermediary representation that enables

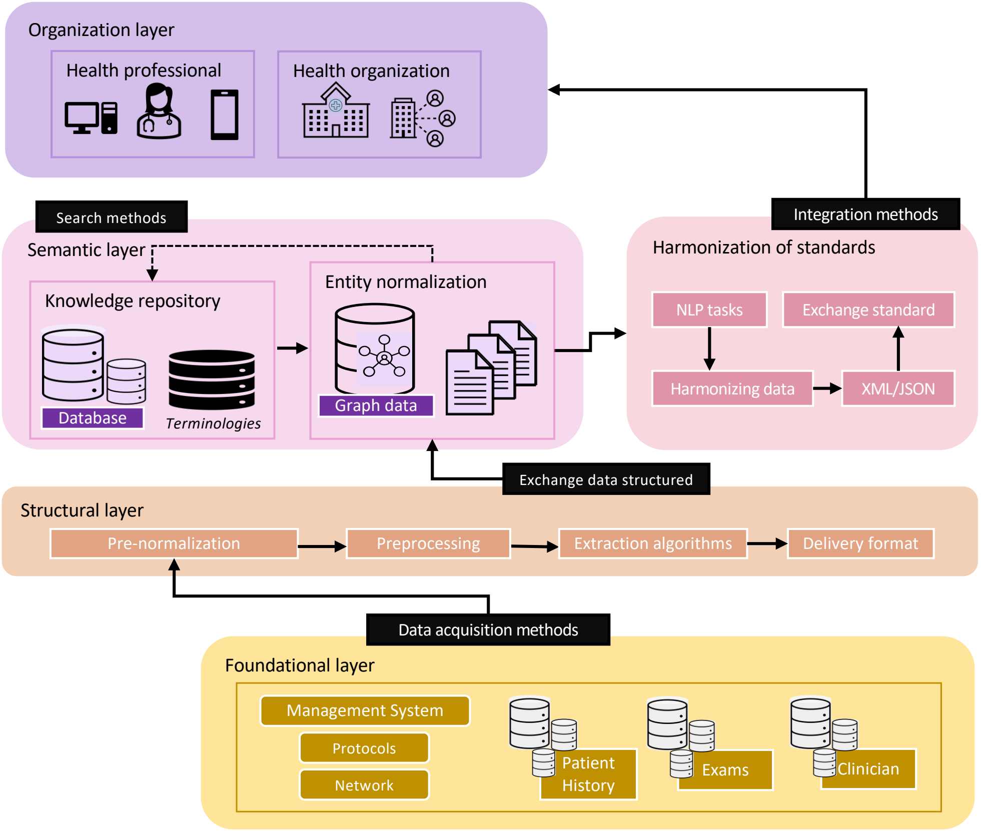
[1] BioPortal CIDO https://bioportal.bioontology.org/ontologies/CIDO/

[2] BioPortal CODO https://bioportal.bioontology.org/ontologies/CODO/

[3] BioPortal COVID-0 https://bioportal.bioontology.org/ontologies/COVIDO/

[4] Complete results can be found in the doctoral thesis available on https://repositorio.jesuita.org.br/handle/UNISINOS/13608

Figure 1. The proposed model to achieve semantic-level data interoperability in non-structured data.



Created by the author.

flexible mapping to different healthcare standards.

The model was evaluated in a real-world case study using COVID-19 clinical data from hospital partners. The proposed model remains independent of technological choices, following best practices in interoperability [Benson and Grieve 2021]. The methodology includes: Dataset development, defining essential medical concepts and annotations. NER and entity normalization, ensuring alignment with structured vocabularies [Zha et al. 2024]. Specialist validation, ensuring extracted data supports clinical decision-making. By integrating NLP and ontologies, the model bridges unstructured narratives and structured health data, enhancing interoperability across healthcare systems. The next section details its architectural design and practical applications in real-world healthcare interoperability scenarios.

4. DISCUSSION

The COVID-19 pandemic exposed critical gaps in healthcare interoperability, emphasizing the need for structured clinical data processing [Raza and Schwartz 2023]. Healthcare institutions faced challenges in standardizing records, highlighting the importance of automated data extraction for patient monitoring. Our research addresses this by proposing a semantic interoperability model that structures free-text clinical data, ensuring seamless integration and information exchange. Brazil's Unified Health System (SUS) decentralization results in heterogeneous EHR formats, complicating integration [Paim 2018]. The National Healthcare Data Network (RNDS) aims to unify data, yet diverse information sources present challenges. Our model mitigates these by harmonizing extracted data with HL7 FHIR and openEHR, bridging gaps in clinical data standardization. To validate our approach, we developed a real-world dataset using COVID-19 clinical notes, ensuring alignment with SOAP documentation standards [Podder et al. 2021]. Given the lack of annotated corpora in Portuguese, we adopted a hybrid annotation strategy [Zha et al. 2024], improving entity recognition and normalization. Beyond entity extraction, our model ensures semantic interoperability by aligning structured data with biomedical ontologies such as COVID-19 Ontology, CIDO, and CODO, facilitating integration into diverse health IT ecosystems. While tested in a COVID-19 case study, its scalability allows applications in oncology, radiology, and primary care. Future work will enhance semantic disambiguation, multilingual capabilities, and adaptation to nursing and administrative records. By structuring unstructured healthcare data, our approach strengthens interoperability, supporting data-driven decision-making and continuity of care across decentralized health systems.

5. CONCLUSION

This research addressed the question: How to extract, disambiguate, and represent unstructured health data for interoperability using hybrid NLP and machine learning approaches? A systematic review in Semantic Interoperability [Mello et al.] identified key gaps in standardizing unstructured clinical data, highlighting challenges and, structuring data across healthcare systems. To address these issues, we proposed a semantic interoperability model for Portuguese clinical notes, integrating NLP and ML for automatic entity extraction and structuring. This approach mitigates the absence of standardized healthcare terminologies, reducing manual effort and supporting clinical decision-making. The model organizes extracted entities into a formal ontology, ensuring alignment with HL7 FHIR and openEHR while maintaining modularity. Additionally, a lexical normalization dictionary enhances data consistency, providing a reusable resource for interoperability. In conclusion, our model proposes structuring clinical notes, promoting standardization, and seamless data exchange. Combining information extraction, disambiguation, and ontological representation strengthens NLP-driven digital health solutions, bridging semantic interoperability gaps.

References

Alexopoulos, P. (2020). *Semantic Modeling for Data*. O'Reilly Media, 1st ed. edition.

Benson, T. and Grieve, G. (2021). Why interoperability is hard. In *Principles of Health Interoperability*, pages 21–40. Springer.

da Silva Jr, J. B., Lima, N. T., Garcia-Saisó, S., Fitzgerald, J., Bascolo, E., Gross Galiano, S., Solis Ortega, A. E., Morales, C., Marti, M., Estela Haddad, A., et al. (2024). Towards 2030: ministerial agreements on information systems and digital transformation for resilient health systems.

El Kah, A. and Zeroual, I. (2021). A review on applied natural language processing to electronic health records. In *2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, pages 1–6. IEEE.

Hasan, S. A. and Farri, O. (2019). Clinical natural language processing with deep learning. In *Data science for healthcare*, pages 147–171. Springer.

HIMSS (2021). *Healthcare Information and Management Systems Society*.

ISO18308 (2011). ISO 18308:2011(en), Health informatics — Requirements for an Electronic Health Record Architecture.

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., et al. (2021). Neural natural language processing for unstructured data in electronic health records: a review. *arXiv preprint arXiv:2107.02975*.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Martin-Sanchez, F. and Verspoor, K. (2014). Big data in medicine is driving big changes. *Yearbook of medical informatics*, 23(01):14–20.

Mello, B., Jos, S., Andr, C., and Carine, L. Use of semantic interoperability standards in Health Records : a systematic review.

Mougin, F., Hollis, K. F., and Soualmia, L. F. (2022). Inclusive digital health. *Yearbook of Medical Informatics*, 31(01):002–006.

Paim, J. S. (2018). Thirty years of the unified health system (sus). *Ciência & Saúde Coletiva*, 23:1723–1728.

Podder, V., Lew, V., and Ghassemzadeh, S. (2021). Soap notes.[updated 2021 sep 2]. *StatPearls [Internet]. StatPearls Publishing. Available from: https://www. ncbi. nlm. nih. gov/books/NBK482263*.

Raza, S. and Schwartz, B. (2023). Entity and relation extraction from clinical case reports of covid-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*, 23(1):20.

Shen, Y.-C., Hsia, T.-C., and Hsu, C.-H. (2021). Analysis of electronic health records based on deep learning with natural language processing. *Arabian Journal for Science and Engineering*, pages 1–11.

Sheth, A. P. (1999). Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In *Interoperating geographic information systems*, pages 5–29. Springer.

Sim, J.-a., Huang, X., Horan, M. R., Stewart, C. M., Robison, L. L., Hudson, M. M., Baker, J. N., and Huang, I.-C. (2023). Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: A systematic review. *Artificial intelligence in medicine*, page 102701.

Sivarethinamohan, R., Sujatha, S., and Biswas, P. (2021). Envisioning the potential of natural language processing (nlp) in health care management. In *2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC)*, pages 189–193. IEEE.

Zha, Y., Ke, Y., Hu, X., and Xiong, C. (2024). Ontology attention layer for medical named entity recognition. *Applied Sciences*, 14(1):421.