LRMoo as the Conceptual Model for the Lem Knowledge Graph

Luiz do Valle Miranda[1,*], Jakub Gomułka[2], Szymon Piotr Kukulak[2], Krzysztof Kutt[1] and Grzegorz J. Nalepa[1]

[1]*Jagiellonian Human-Centered AI Lab, Mark Kac Center for Complex Systems Research, Institute of Applied Computer Science, Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, prof. Stanisława Łojasiewicza 11, 30-348 Kraków, Poland*

[2]*Faculty of Humanities, AGH University of Kraków, Czarnowiejska 36, 30-054, Kraków, Poland.*

Abstract

This paper explores the challenges of representing the works of Stanisław Lem (1921–2006) in a knowledge graph (KG), focusing on their complex versions across languages, editions, and collections. Lem's works, characterized by an evolving content and multiple versions, require a refined approach to capture their nuances. We propose using LRMoo, an extension of CIDOC-CRM, to model Lem's prose. Through a case study of *The Star Diaries*, we demonstrate how LRMoo can appropriately represent these complexities.

Keywords

LRMoo, Cultural Heritage, Stanisław Lem, Linked Data, Knowledge Graph

1. Introduction

Stanisław Lem (1921–2006) is a well-known Polish writer whose works have been translated into over fifty languages. He is famous for his numerous science fiction novels and stories—including *Solaris*, *The Invincible*, *His Master's Voice*, *Star Diaries*, and *Golem XIV*—as well as for his non-fiction works such as *Dialogs*, *Summa Technologiae*, *The Philosophy of Chance*, and *Science Fiction and Futurology*. The total number of worldwide editions of his books—including both standalone novels, essays, and various collections of short stories—has exceeded 1,000. Lem was one of those authors who, during his lifetime, prepared multiple versions of some of his works, removing old chapters or stories while also adding new ones. At times, he instructed translators to shorten certain pieces. As a result, some of his works and collections exist in multiple versions within the same language, and some stories and essays differ in content across different translations. [1] refers to this differing aspect of Lem's prose as a "book in motion".

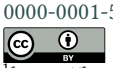
There have been different ways that Stanislaw Lem's prose has been cataloged and made available. A recent publication by Victor Yaznevich [2] covers most of the writer's non-fiction works. There are also a few other initiatives collecting information about Lem's works published around the world, including The Lem Encyclopedia in a form of a wiki site[1] and a web archive on Lem's world bibliography run by Vladimir Borisov in Russian[2]. All available databases are created as a hobby by enthusiasts of Lem's work, and none of them contain a complete catalog of his works and editions. This led to the idea of creating a knowledge base on Lem's work that would collect, integrate, and share information from all possible sources. The project is affiliated with a public institution, the Jagiellonian University Museum, and is being developed using widely recognized standards for this type of knowledge base.

SemDH 2025: Second International Workshop of Semantic Digital Humanities. Co-located with ESWC 2025, June 02, 2025, Portoroz, Slovenia.

*Corresponding author.

luiz.dovallemiranda@doctoral.uj.edu.pl (L. d. V. Miranda); jgomulka@agh.edu.pl (J. Gomułka); skukulak@agh.edu.pl (S. P. Kukulak); krzysztof.kutt@uj.edu.pl (K. Kutt); gjn@gjn.re (G. J. Nalepa)

0000-0003-1838-5693 (L. d. V. Miranda); 0000-0002-9100-0334 (J. Gomułka); 0009-0009-2701-6567 (S. P. Kukulak); 0000-0001-5453-9763 (K. Kutt); 0000-0002-8182-4225 (G. J. Nalepa)



[1] https://lem.wiki/w/The_Lem_Encyclopedia

[2] https://bvi.rusf.ru/lem/lem.htm

This initiative is the Lem Knowledge Graph (LKG), which is being developed as part of the Jagiellonian University's flagship project, CHExRISH.

The aim of the CHExRISH project is to integrate cultural heritage (CH) resources from different units of the Jagiellonian University as linked data (LD) and to present a prototype of a semantic portal, called the Jagiellonian University Heritage Metadata Portal (JUHMP). By modeling cultural heritage resources as LD, galleries, libraries, archives, and museums (GLAM) aim at leveraging the technology of Knowledge Graphs (KG) and the principles of the Semantic Web towards the aggregation of data from different institutions, browsing and search through a semantically meaningful representation of a certain domain, and the use of visualization and knowledge discovery tools to assist digital humanities research [3]. LKG is planned as an extension to JUHMP, but that can also work as a standalone knowledge base.

The basis for a semantic portal is the representation of data using controlled vocabularies and a shared conceptualization of a certain domain. An important decision in a CH-LD project is what standards to follow to maximize preservation, interoperability and dissemination of knowledge relating, among these the choice of an ontology to map domain knowledge to a shared representation [4]. There are a variety of existing ontologies to model CH resources, among others the Europeana Data Model[3] (EDM) [5], which provides a framework for the description of (digitalized) cultural heritage objects to be easily aggregated into the Europeana portal, the Records in Contexts Ontology[4] (RiC-O) [6], which enables the representation of archival information within a broader context, the Lightweight Information Describing Objects[5] (LIDO) [7], which supports the exchange of metadata about museum collections, and CIDOC-CRM[6] [8], an international standard (ISO 21127:2014) for cultural heritage data modeling and sharing across different institutions. Since the development of LKG is part of the more extensive CHExRISH project which already uses CIDOC-CRM as the conceptual model for LD to be integrated into JUHMP, LKG has to follow a model at least compatible with CIDOC-CRM to ensure consistency and interoperability across the different components of the project.

Representing Lem's prose with CIDOC-CRM base ontology is possible, however, such a conceptual model flatten nuances in the way Lem's works diverge and connect in the levels of stories, collections, languages and publications. As will be described in detail, Lem's prose—for a number of reasons—requires a more extensive method of description, in order to properly reflect all the internal relations. LRMoo[7], an extension to CIDOC-CRM, appears to be an adequate alternative for the conceptual model of the LKG. LRMoo is designed to offer a more refined description of bibliographic information, maintaining the semantics of different creative aspects of a literary work.

In this paper, we aim at presenting the work in progress in modeling Lem's fiction and non-fiction as a KG based on the LRMoo conceptual model. First, we want to explain the particularities of Lem's prose, and exemplify how the works differ according to language or collection. This effort will give the requirements to develop a conceptually appropriate LKG. Second, we want to present work in progress of actually mapping LRMoo classes and properties to Lem's publication information. We present the example of Lem's *Star Diaries* and its components. We conclude the paper presenting future directions for the construction of LKG.

2. Lem's prose and the requirements for LKG

A complexity of internal relations links different pieces of Lem's prose on various levels in both his fiction and non-fiction. This interconnectedness of literary entities is not entirely unique to Lem's work. To some extent, a similar fix-up approach can be observed in the prose of several American authors from the Golden Age of science fiction ([9], pp. 307-311). However, Lem's practice of reusing the same texts multiple times and altering the content of the same books in subsequent editions makes the issue far more pronounced. As an example, Jerzy Jarzebski [1], referring to *Summa Technologiae*, described

[3] https://pro.europeana.eu/page/edm-documentation
[4] https://www.ica.org/resource/records-in-contexts-ontology/
[5] https://cidoc.mini.icom.museum/working-groups/lido/
[6] https://cidoc-crm.org/
[7] https://cidoc-crm.org/lrmoo

this aspect of Lem's prose as a "book in motion", a term that seems fitting for many of Lem's works. To successfully model the relationships within Lem's writings, it is first necessary to understand this fluidity. Examining key instances of the evolution of his work will help define the requirements for accurately modeling both fiction and non-fiction in LKG.

2.1. The requirements related to Lem's fiction

A practice analogous to the fix-up technique in Lem's prose is the reusing of the short story 'Kryształowa kula' (A Crystal Ball), originally published in 1954 in the collection *Sezam i inne opowiadania* (Sesame and Other Stories). Thirty years later, a heavily reworked portion of it appeared as part of the novel *Fiasco*. A somewhat similar situation occurred with *The Invincible*, which was first published as part of the collection *Niezwyciężony i inne opowiadania* (The invincible and Other Stories) in 1964 and later came to function as a standalone novel.

Short stories in Lem's prose often appear in multiple collections. In most cases, they were first published alongside unrelated works from the same period and later reorganized into thematically cohesive collections. For example, the first stories about Ijon Tichy – among others, 'The Twenty-second Voyage', 'The Twenty-third Voyage' – were published as early as 1954, while their dedicated collection, *The Star Diaries*, was first compiled in 1957. Furthermore, some thematically cohesive collections do not necessarily include all stories belonging to a given cycle, as seen in the omission of 'The Twenty-sixth and Last Voyage' in *The Star Diaries*. Additionally, certain translations omitted portions of the source material, often deliberately altering their tone, as in the case of 'The Twenty-second Voyage'.

Memoirs of a Space Traveller appeared in English translation with the addition of two short stories ('The Eighteenth Voyage' and 'The Twenty-fourth Voyage'), which were not included in the English edition of *The Star Diaries* despite being part of the original Polish collection. Several other short stories that shifted between the two collections in different Polish editions do not appear in either of their English counterparts.

2.2. The requirements related to Lem's non-fiction

Similar phenomena can be observed in Lem's non-fiction works. One key issue is the addition and removal of chapters in different editions of the same book while the title remains unchanged. As a result, a single work may exist in multiple original versions, each containing different subsets of essays, some of which republished in other collections. For instance, *The Philosophy of Chance* underwent several revisions, with Lem removing most chapters devoted to the critique of structuralism.

Another issue is the nature of posthumous collections, which generally consist of older texts with only a few entirely new additions, such as Polish translations of Lem's essays originally published in other languages, as seen in *Diabeł i arcydzieło* (Devil and the Masterpiece). However, these works are often restructured to form new, cohesive entities. Notably, Lem employed a similar approach during his lifetime. For example, he compiled *Rozprawy i szkice* (Essays and Sketches), later republished as *Mój pogląd na literaturę* (My View on Literature), primarily from his prefaces to science fiction novels, mostly by Western authors. In the revised structure, these prefaces became part of Lem's broader reflection on literary issues.

Taking these observations into consideration, it becomes clear that only a complex, multilayered conceptual model can faithfully capture the intricacies of the evolution of Lem's oeuvre. An ontology for LKG should, therefore, distinguish between the main thread that unites different versions of a story and their actual editions. Furthermore, it should account for the evolving composition of collections. Another factor to be considered is the fluid boundary between a short story and a novel, as rewritings can easily shift a work between these two categories. Finally, the model should capture the distinctiveness of different versions of Lem's work based on the language of publication.

3. LRMoo as the basis for LKG

LRMoo is an officially recognised extension of CIDOC-CRM designed to support richer semantic relationships for bibliographic data. Among the implementations of this model in linked open data cultural heritage projects, in [10] LRMoo has served as the basis for a knowledge base on Medieval and Renaissance geographic Latin literature to be used as the backbone of a user-friendly web application for access and visualization of bibliographic data. [11] presents a discussion and a case study on how finding aids can be enriched by drawing upon the complex collections of properties that compose LRMoo offering new possibilities of archival and textual research. Similarly to these cases, we leverage the intricacies of LRMoo to construct a comprehensive semantic representation of Lem's works, thus facilitating new possibilities for finding, visualizing and researching his prose.

LRMoo is structured around four core classes – *Work*, *Expression*, *Manifestation*, and *Item* – each corresponding to a distinct level of representation of a literary work. At each of these levels, the ontology supports the formation of hierarchical structures, allowing the specification of superordinate entities (e.g., novels, collections of stories) as well as subordinate entities (e.g., chapters, individual stories). This hierarchical differentiation between conceptual layers of literary representation is what establishes LRMoo as an appropriate base ontology for LKG.

The highest level of abstraction in LRMoo is represented by the class *Work*, which is intended to group "distinct intellectual ideas conveyed in artistic and intellectual creations, such as poems, stories, or musical compositions" ([12], p. 23). At this level, we assume that we are dealing with the concept of a work, which can have multiple realizations, such as different versions or translations. *Work* is a subclass of CIDOC-CRM class *Propositional Object*.

At a lower level of abstraction, the class *Expression* is defined, whose entities are linked to those of *Work* through the properties *is realised in* (*realises*). *Expression* thus encompasses the concretisations of *Work* entities "in the form of identifiable immaterial objects" ([12], p. 24). This implies that different versions of a work, including its translations into various languages, are classified under this class. The *Expression* class is a subclass of the CIDOC-CRM class *Information Object*.

At the next level of concretization, we find entities of the class *Manifestation*. At this stage, not only the form of the work as a specific sequence of textual or graphical symbols is relevant, but also "the manner in which they are presented to be consumed by users, including the kind of media adopted" ([12], p. 25). In other words, this class includes specific editions of a given work, and its entities are related to entities of *Expression* by the property *is embodied in* (*embodies*). Like *Expression*, *Manifestation* is also a subclass of CIDOC-CRM class *Information Object*.

In LKG, Lem's works will be represented at multiple levels—potentially even all four. At the *Work* level, his writings will be treated as intellectual constructs. As shown in Fig. 1, *Work* representations may refer to both the ideas of collections—such as *The Star Diaries*, *The Memoirs of a Space Traveler*, and *Sesame and Other Stories*—as well as to the ideas of individual short stories, such as *The Twenty-first Voyage* and *The Twenty-second Voyage*. The relation *forms part of* will allow for indicating hierarchical order between these two groups. Since the idea of a collection does not have a fixed structure, any reason for subsuming a piece into a particular collection will be reflected through the relation linking both representations. As a result, a single short story may be included in multiple collections.

At the *Expression* level, these intellectual constructs will be realized in specific ways. This will be captured by the relation *is realised in*. For instance, the *Work* instance *The Star Diaries* will be realized by several *Expression* instances, including *The Star Diaries v1 EN*, which will represent the first structured version of the collection in English translation. This version will not be identical to *The Star Diaries v1 PL*, as it will have a different list of components. In turn, *The Twenty-second Voyage v1 EN* will represent the English version of the short story with its textual content determined, while *The Twenty-second Voyage v1 PL* will represent the original Polish version. The relation *is derivative of* will be used to link the two.

Expression instances will be connected to specific editions, which will belong to the *Manifestation* class. For example, the 1976 New York English edition of *The Star Diaries* will be modeled as an embodiment of *The Star Diaries v1 EN*. A part of it, *The Twenty-second Voyage*, will be represented as the *Manifestation*

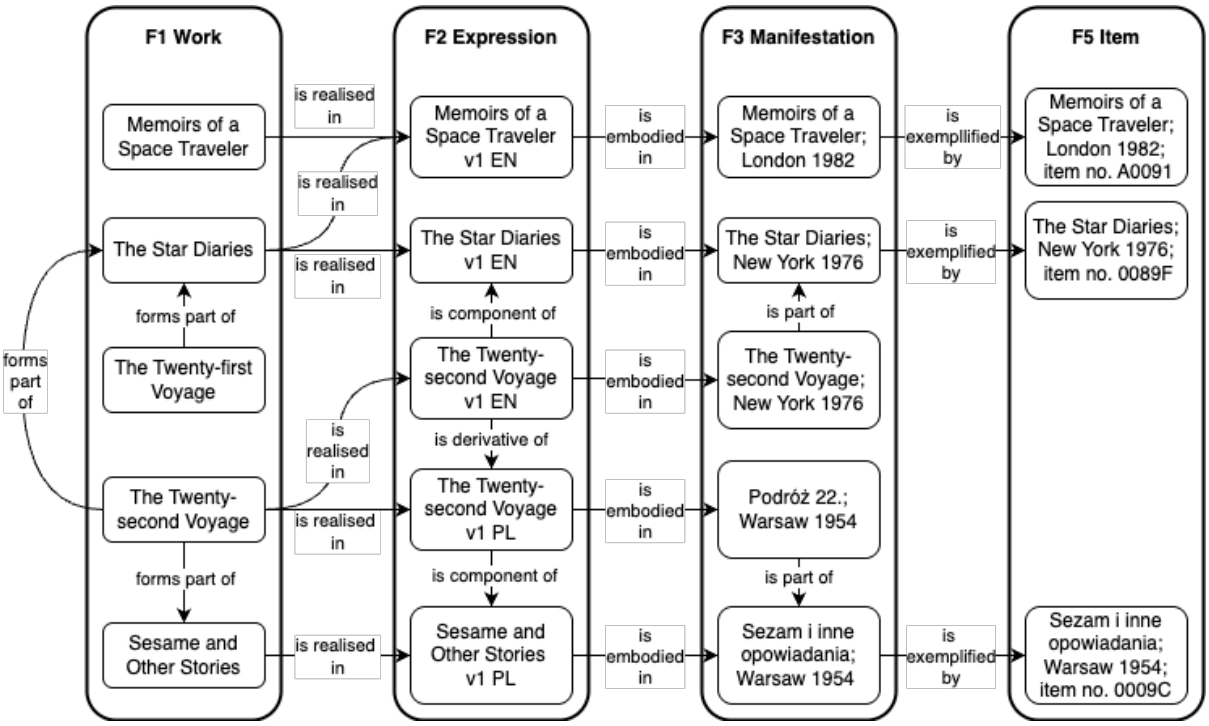


Figure 1: LRMoo multilayered modeling of *The Star Diaries* and some of the short stories that compose it

instance *The Twenty-second Voyage; New York 1976*, which will embody *The Twenty-second Voyage v1 EN*. The *Manifestation* level will be specific enough for its instances to be named according to the actual titles of the editions, such as *Sezam i inne opowiadania; Warsaw 1954* as the embodiment of *Sesame and Other Stories v1 PL*.

Individual copies of specific editions will be represented as *Item* instances. If a particular copy is of interest (e.g., it contains handwritten notes), it will be modeled as the exemplar *The Star Diaries; New York 1976; item no. 0089F*. These will be related to their corresponding *Manifestation* instances by the relation *is exemplified by*.

There remains a problem to be addressed in representing Lem's works within the LRMoo ontology. As noted in Section 2.1, the English translation of *The Twenty-second Voyage* differs in content from the Polish original. This is a significant feature that should be represented in the graph. LRMoo does not offer a straightforward solution to this issue: the relation *is derivative of* merely indicates a translation relationship and does not express the nature or degree of variation. LRMoo, however, supports the reification of the property *is derivative of* with the goal of connecting it to an instance of the CIDOC-CRM class *type*. Another solution is to supplement the ontology with sub-properties for *is derivative of*, such as *is shortened version of*, *is extended version of*, and so on. At the current moment, no final decision has yet been reached regarding this modeling challenge.

4. Conclusion and further works

The case of modeling *The Star Diaries* as an intellectual idea, a certain version, an edition and a physical item, together with the respective modeling of some of its composing short-stories has shown that LRMoo is effective in representing relationships between various publications of Lem's works, including cases in which texts are repurposed, altered, and reorganized in diverse ways depending on the edition and language. Despite the adequate expressive power of LRMoo, there are still some factors to consider for future work in LKG.

As part of the CHExRISH project, LKG's knowledge base is expected to be included in the JUHMP portal. However, the knowledge base powering the latter is CIDOC-CRM-based, thus it is not compatible

with the nuances of LRMoo. A future challenge, thus, is to develop a flattening mapping between LRMoo and CIDOC-CRM in a way to present, despite the lack of nuance, appropriate and adequate bibliographic information about Lem's publications.

Furthermore, the LKG project aims to expand the knowledge graph to include information on entities beyond publications, such as letters. Neither CIDOC-CRM nor LRMoo inherently support the modeling of concepts critical for representing epistolary resources. For instance, terms like "sender" and "recipient" are not included within the classes or properties of these ontologies. Moreover, to date, there is no official extension of CIDOC-CRM specifically designed to model letters and their epistolary relationships[8]. In the case of the development of such an epistolary extension of CIDOC-CRM, there is still the challenge of harmonizing such an extension with LRMoo.

Finally, a semantic portal independent of JUHMP is planned to interactively present data from LKG. Digital showcases could potentially be prepared on varying aspects of Lem's work, for example, repeating themes and ideas on different phases of the author's oeuvre, or a list of the fictional characters he created.

Declaration on Generative AI

This publication benefited from the use of language models—gpt-4o, gpt-4o-mini and DeepSeek-V3—to support proofreading and enhance readability. All generated text was reviewed and edited, and the authors take full responsibility for the publication's content.

Acknowledgments

This publication was funded by a flagship project "CHExRISH: Cultural Heritage Exploration and Retrieval with Intelligent Systems at Jagiellonian University" under the Strategic Programme Excellence Initiative at Jagiellonian University.

The research for this publication has been supported by a grant from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University.

References

[1] J. Jarzębski, "summa technologiae" i jej potomstwo, in: Summa technologiae, Wydawnictwo Literackie, Kraków, 2000.

[2] V. Yaznevich, Stanislaw Lem Non Fiction. Bibliography and Personalities, Medisont, Minsk, 2021.

[3] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187 – 193. doi:10.3233/SW-190386.

[4] W. Z. Alma'aitah, A. Z. Talib, M. A. Osman, Opportunities and challenges in enhancing access to metadata of cultural heritage collections: a survey, Artificial Intelligence Review 53 (2020) 3621–3646. URL: https://doi.org/10.1007/s10462-019-09773-w. doi:10.1007/s10462-019-09773-w.

[5] A. Isaac, Europeana Data Model Primer, Technical Report, Europeana, 2013. https://pro.europeana.eu/page/edm-documentation.

[6] D. Pitti, G. McCarthy, B.-F. Popovici, Records in Contexts (RiC). An Archival Description Draft Standard, Technical Report, ICA Experts Group on Archival Description, 2017. https://www.alaarchivos.org/wp-content/uploads/2018/01/1.-Daniel-V.-Pitti.pdf.

[7] J. Lindenthal, H.-L. Meiners, D. Balzer, LIDO Primer, Technical Report, CIDOC LIDO Working Group, 2023. https://lido-schema.org/documents/primer/2023-09-20/lido-primer.html.

[8] D. Oldman, The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER, Technical Report, CRM Labs, 2014. http://old.cidoc-crm.org/docs/CRMPrimer_v1.1.pdf.

[8] [13] presents a non-official extension of CIDOC-CRM to capture semantic aspects of correspondences

[9] I. Asimov, Serials, in: Gold: The Final Science Fiction Collection, Eos, New York, 2003.
[10] V. Bartalesi, D. Metilli, N. Pratelli, P. Pontari, Towards a knowledge base of medieval and renaissance geographical latin works: The imago ontology, Digital Scholarship in the Humanities 37 (2021) 34–50. URL: https://doi.org/10.1093/llc/fqab060. doi:10.1093/llc/fqab060. arXiv:https://academic.oup.com/dsh/article-pdf/37/1/34/42986942/fqab060.pdf.
[11] F. Giovannetti, F. Tomasi, The linked finding aid as a platform for textual research: The case study of the giuseppe raimondi archive, in: L. C.T, R. C, N. F, R. I, F. N (Eds.), CEUR Workshop Proceedings, volume 3019, CEUR-WS, 2021, pp. 104 – 113. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121099627&partnerID=40&md5=433315797cb5a4007a01fc8f0d90584f, iSSN: 16130073 Type: Conference paper.
[12] I. L. W. Group, Lrmoo: Object-oriented definition and mapping from the ifla library reference model, 2024. URL: https://cidoc-crm.org/sites/default/files/LRMoo_V1.0.pdf, last accessed: 12 March 2025, CIDOC CRM.
[13] A. Deicke, C. Seibold, E. S. Cronauer, Correspondence and epistolary research ontology, 2024. URL: https://lod.academy/cer/vocab/ontology/, accessed: 2025-03-14.