How Wikidata and Wikipedia can help identify, improve and diversify the world literary canon in a multilingual context

Tomás Saorín (User:Tsaorin) Juan-Antonio Pastor-Sánchez (User:Japastorsanchez) Mª José Baños-Moreno (User:Jojo.alguazas2)
Department of Library and Information Science. University of Murcia (Spain)

Abstract

In Wikipedia and Wikidata, valuable data for studying the impact of literary works, especially narratives like novels, books, sagas, or tales, exists. This proposal aims to refine the taxonomy of literary and printed works through automated analysis in Wikidata. Apply a global relevance metric to identify prominent works across languages and time. Compare literature coverage across Wikipedias to reveal cultural proximities, gaps, and improvement needs. Provide a website for visualizing results, utilizing a public dataset for replicable research, with an automated data update mechanism.

Introduction

This proposal is based on a very simple working hypothesis: **Could we use Wikidata and Wikipedia as a source to identify a global literary canon to foster quality and coverage improving across different languages and communities?** The concept of a universal literary canon has been subject of relevant debates and it's. However a contented issue, it still worths attention from the critical cultural studies. Some of these criteria or canonicity are based on aesthetic, historical, cultural, or ideological values, while others are based on quantitative measures, such as popularity, influence, or recognition. In this proposal, we adopt a quantitative approach, using the data from Wikidata and Wikipedia as sources of information about the literary works and their reception. We assume that these data reflect the collective judgment of the Wikimedia community, mixing specialized and the general opinions, and that they can provide insights into the global and multilingual dimensions of the literary canon.

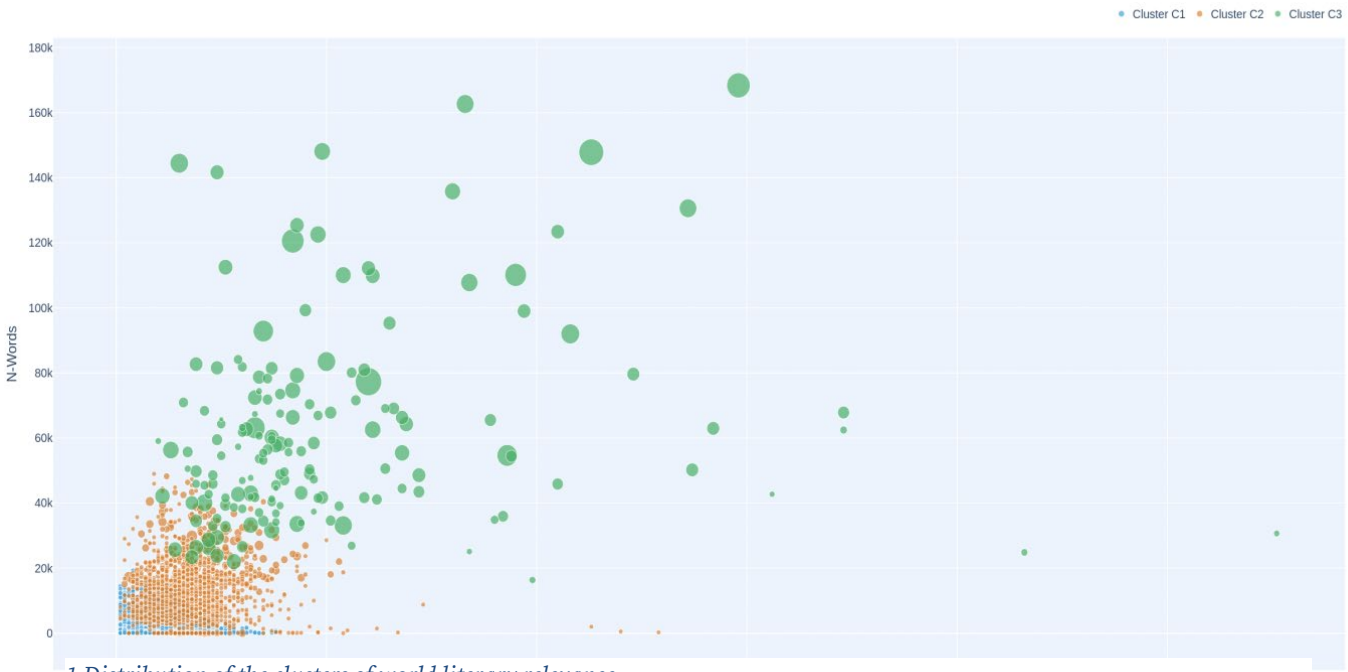
This proposal aims to develop a method to fix the classification of literary items in Wikidata and delimitate all data related to literary works. We will select a subset of relevant books from the point of view of world literatures and measure the presence of literatures in one specific language in the editions of Wikipedia in other languages to map the proximities between cultural communities. We will also improve the method to score aggregated creative works.

This is an ongoing research within a broader project about cultural artifacts rankings. The

milestones proposed start at **June 2024** with the launch of a website about literary categories and clasess in Wikipedia/Wikidata, that will explain the different issues to outline the literary field. At **fall 2024** will be launched a website about the scoring and ranking tests accomplished on literary works data and contents (Wiki3DRank website). Complementary, at **2025 spring** will be launched the website that maps world literatures on each Wikipedia languages. Alongside this period, research results will be be advanced in our website “Fictional Metadata”, both in english and spanish (https://www.um.es/metadatosdeficcion) and also discussed and presented at research seminars, meetings and conferences.

Related work

There are thrilling Wikiprojects at Wikidata and Wikipedia about Books, Novels, Fictional Universes and others. Previous research has showed that information about authors is more detalailed and extent that the one for their works, content, characters or, in a broad sense, fictional universes. Our previous 2023 research paper "A universal literary canon based on multilingual encyclopedic data: Proposal of a method for the ranking of literary works using quantitative data obtained from Wikidata and Wikipedia" provides the foundation for this proposal. The most relevant works in this field include those of the *Wikipedia Diversity Observatory*, a project aimed at understanding and increasing diversity within Wikipedia content and communities (https://meta.wikimedia.org/wiki/WDO), and also those included in the recent monograph issue focused on Wikipedia, Wikidata, and World of the Literature of the *Journal of Cultural Analytics* (https://culturalanalytics.org/issue/7259). Also, *WikiRank project* (https://www.wikirank.net) is an initiative that aims to measure the quality and popularity of Wikipedia articles in different languages. It uses various indicators, such as



1 Distribution of the clusters of world literary relevance

references, readability, completeness, and citations, to rank the articles according to their reliability and relevance.

Methods

The methods used in this proposal will include:

- Processing and automatic analysis of taxonomies using contextual content processing tools.
- Improvement and enrichment of data related to selected works in different relevant proposals for literary canon by means of media, institutions, awards, and interest groups.
- Optimization of SPARQL queries on the Wikidata Query Service.
- Statistical validation of the minimum set of variables presents in literary works in Wikipedia and Wikidata to obtain a representative global ranking.

Expected output

The anticipated outputs of the project encompass the following elements:

1. Comparative World Literature website:

- Development of a Comparative World Literature website, considering the relevance of each literary work based on the calculated rank metric in the project and its dissemination across languages.
- Primary intended audience: Scholars, students, and enthusiasts of world literature.

2. Temporal distribution of works in Spanish (top) and English (bottom). Source: own preparation.

- Benefit: Access to a comprehensive platform for exploring and comparing literary works from diverse cultural and linguistic backgrounds, facilitating in-depth analysis and cross-cultural understanding.

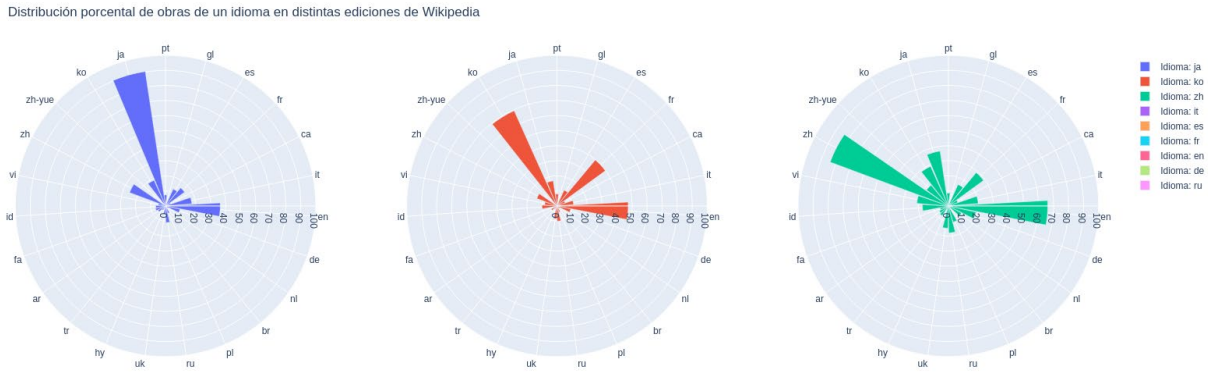
2. **Wikipedia article tracking:**
   - Integration in the website of a system to track pending Wikipedia articles for creation or improvement in various languages, focusing on significant works within the global literary canon.
   - Primary intended audience: Wikipedia editors, language enthusiasts, and researchers.
   - Benefit: Facilitation of collaborative efforts to enhance the representation of globally significant gaps and balances between different language editions of Wikipedia, promoting cultural exchange and knowledge dissemination.

3. Literary data curation dashboard:
   - Provision of indicators for data completeness pertaining to literary works and the establishment of a coherent and organized taxonomy of classes for types, genres, and artistic forms.
   - Primary intended audience: Researchers, librarians, and data curators.
   - Benefit: Enablement of efficient data assessment and navigation through a structured taxonomy, enhancing the accessibility and usability of multilingual encyclopedic data related to literary works.

These outputs aim to contribute to the accessibility, understanding, and preservation of global literary heritage, fostering cross-cultural dialogue and scholarly engagement within the field of world literature.

Risks

The project minimizes risks by leveraging existing Wikimedia ecosystem data. Enrichment sources are accessible, and proof of concept tests have been conducted. The main research team has a stable academic connection, and support staff efforts for development, visualization, and data exploitation are reasonable for expected results..



3. Japanese literature of JA-WIKI present in other Wikipedias

Community impact plan

1. Cultural Collaborations: Collaborating with cultural organizations to host public events, including literary festivals and workshops, fostering awareness and appreciation for global literary heritage.
2. Wikimedia Partnership: Partnering with Wikimedia volunteer editors to enrich Wikipedia's representation of global literary works, encouraging multilingual content creation and enhancement.
3. Multilingual Outreach: Implementing strategic multilingual communication to engage diverse language communities, promoting the widespread adoption and use of project outputs globally.

Evaluation

1. User engagement metrics: Tracking website traffic, user feedback, and platform engagement to measure audience interest and utilization.

2. Wikimedia content contributions: Monitoring contributions to Wikimedia platforms to gauge project influence within the Wikimedia community.

3. Publication impact: Evaluating dissemination in international journals, including citation metrics and academic feedback, to measure scholarly impact.

Budget

The proposed budget for the project includes the following items:

1. Web design and visualization: 2500$
 - Design and development of the Comparative World Literature website, including user interface design, data visualization, and website hosting.
 - Cost breakdown: Web design and development (2500$), website hosting (Institutional University of Murcia).

2. Workshop at Wikimania or Wikimedia Community International Meeting: 1600$
 - Conducting an event to facilitate knowledge exchange and collaboration among editors and researchers. The workshop aims to enhance the skills and expertise of participants in the Wikimedia community.
 - Cost breakdown: Travel Expenses for Invited Editors/Researchers: (1400$), Administrative costs (200$).

3. Open access publication costs: 500$
 - Covering the costs of publishing project results in open access international journals.
 - Cost breakdown: Publication fees (400$), administrative costs (100€).

The total proposed budget for the project is 4,600$. The budget items have been carefully selected to ensure the successful implementation of the project and the dissemination of its results to a wide audience.

Prior contributions

Saorín, T., Pastor-Sánchez, J. A. How to end lists of the best books once and for all. The Conversation (Spanish edition), April 2023, https://theconversation.com/como-acabar-de-una-vez-por-todas-con-las-listas-de-los-mejores-libros-203353

Pastor-Sánchez, J. A., Saorín, T. Measuring the literary field and creative works: the case of world literary canon according to Wikipedia and Wikidata. International Workshop Wikipedia, Wikidata and Wikibase: Usage Scenarios for Literary Studies. Frei Universität Berlin, 10-11th October 2023, http://hdl.handle.net/10201/134803

Pastor-Sánchez, J. A., Saorín, T., Baños-Moreno, M.J. (2023). Un canon literario universal basado en datos enciclopédicos multilingües: propuesta de un método de medición de obras literarias usando datos cuantitativos obtenidos de Wikidata y Wikipedia. Revista Española de Documentación Científica, 46 (3), e366. https://doi.org/10.3989/redc.2023.3.2013

Saorín, T. Facing Wikipedia in university teaching and learning: a boost to reshape educational & scientific contents production and publishing. International Seminar Wikipedia and University: Research and teaching experiences, nov, 2023, Internet Interdisciplinary Institute (IN3) UOC, http://hdl.handle.net/10201/36784

References

Algee-Heweitt, M., Allison, S., Gemma, M., Heuser, R., y Moretti, F. (2018). Canon/archivo: dinámicas de largo alcance y campo literario. En F. Moretti (Ed.), Literatura en el laboratorio: canon, archivo y crítica literaria en la era digital, 131-181. Gedisa

Bianchini, C., y Sardo, L. (2022). Wikidata : a new perspective towards universal bibliographic control. JLIS, 13(1). DOI: https://doi.org/10.4403/jlis.it-12725

Hill, B., y Shaw, A. (2020). The Most Important Laboratory for Social Scientific and Computing Research in History. En J. Reagle y J. Koerner (eds.), Wikipedia @ 20: Stories of an Incomplete Revolution. The MIT Press. DOI: https://doi.org/10.7551/mitpress/12366.001.0001

Hube, C., Fischer, F., Jäschke, R., Lauer, G., y Thomsen, M. R. (2017). World Literature According to Wikipedia: Introduction to a DBpedia-Based Framework. arXiv. Disponible en: http://arxiv.org/abs/1701.00991

Jemielniak, D., y Wilamowski, M. (2017). Cultural diversity of quality of information on Wikipedias. Journal of the Association for Information Science and Technology, 68(10), 2460-2470. DOI: https://doi.org/10.1002/asi.23901

Lemus-Rojas, M., y Pintscher, L. (2018). Wikidata and Libraries: Facilitating Open Knowledge. En M. Proffitt (ed.), Leveraging Wikipedia: Connecting Communities of Knowledge, 143-158. IL: ALA Editions. Disponible en: https://scholarworks.iupui.edu/handle/1805/16690

Lewoniewski, W., Węcel, K., y Abramowicz, W. (2019). Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. Computers, 8(3), 60. DOI: https://doi.org/10.3390/computers8030060

Miquel-Ribé, M. (2019). The Sum of Human Knowledge? Not in One Wikipedia Language Edition. Wikipedia@20. Disponible en: https://wikipedia20.mitpress.mit.edu/pub/26ke5md7/release/15

Miquel-Ribé, M., y Laniado, D. (2018). Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. Frontiers in Physics, 6, Article 54. DOI: https://doi.org/10.3389/fphy.2018.00054

Miquel-Ribé, M., y Laniado, D. (2021). The Wikipedia Diversity Observatory: helping communities to bridge content gaps through interactive interfaces. Journal of Internet Services and Applications, 12(1), 10. DOI: https://doi.org/10.1186/s13174-021-00141-y

Piscopo, A., y Simperl, E. (2018). Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. Proceedings of the ACM on Human-Computer Interaction, 2, 1-18. DOI: https://doi.org/10.1145/3274410

Skiena, S. S., y Ward, C. (2014). Who's bigger? Where historical figures really rank. Cambridge University Press. Venuti, L. (2008). Translation, interpretation, canon formation.