

# Emerging Patterns in Romanian Literature and Interactive Visualizations based on the *General Dictionary of Romanian Literature*

Irina Toma, Laurentiu-Marian Neagu, Mihai Dascalu, Ștefan Trăușan-Matu

University Politehnica of Bucharest

313 Splaiul Independentei, 060042, Bucharest, Romania

{irina.toma, laurentiu.neagu, mihai.dascalu, stefan.trausan}@upb.ro

Laurențiu Hanganu, Eugen Simion

The “G. Călinescu” Institute of Literary History and Theory, Romanian Academy

Calea 13 Septembrie, Bucharest, Romania

{institutulcalinescu, eugen.ioan.simion}@gmail.com

DOI:10.37789/rochi.2020.1.1.4

## ABSTRACT

The General Dictionary of Romanian Literature (DGLR) is a comprehensive work carried out by researchers from the literary institute of the Romanian Academy. DGLR offers detailed information about writers, editors, translators, literary publications, and cultural institutions that contributed to the Romanian national literature. The current work presents interactive web visualizations, based on statistical studies performed on the DGLR corpus, such as: biographical information; geographic information, including countries that part of the most important writers have visited, lived in, or studied in; active literary entities per year; timeline of publications for important writers. The purpose of the visualizations is to provide overviews regarding the Romanian literature to the general audience. In addition, our views offer valuable insights on the writers and their work across time. A survey was conducted on 20 users and most of them had a pleasant experience; recommendations on future developments were also provided.

## Author Keywords

Analytical approach; Quantitative study; DGLR; Romanian writer; ReaderBench framework; Interactive visualizations.

## ACM Classification Keywords

H.5.2: Information interfaces and presentation (e.g., HCI): User Interfaces;

I.2.7 Natural Language Processing: Discourse, Language parsing and understanding, Text analysis.

## General Terms

Text analysis

## INTRODUCTION

The ongoing work of the Romanian Academy for the national literature digitalization includes two important projects: a) the General Dictionary of Romanian Literature (DGLR), which contains information on the representative writers and institutions that contributed to national literature, and b) the Chronology of Romanian Literary Life (CVLR), which maps the relevant Romanian literary events between 1944 and 2000. The corresponding works are available through two channels, namely in printed form and in the INTELIT web platform. Several statistical analyses were performed using DGLR and CVLR corpuses, and corresponding results were integrated into the ReaderBench framework<sup>1</sup> [10].

The current work presents interactive web visualizations of writers' statistics based on DGLR, their integration into the ReaderBench platform, together with a qualitative study on the visualizations' usefulness and ease of use. The visualizations provide a broad, general perspective on the Romanian literature, through various charts depicting key points of writers' lives and their writings.

The paper is structured as follows. The next section presents the state-of-the-art, highlighting similar studies and available types of visualizations. Next, the third section presents the corpus, together with data processing techniques and technologies used for storage, integration, and visualizations. The fourth section presents interpretations of the views, followed by an evaluation based on a questionnaire. The last section draws the

---

<sup>1</sup> <http://readerbench.com/>

conclusions from the current analysis and outlines possible directions of future development.

### STATE OF THE ART

Most analyses on literature are based on analytical approaches. For example, Moretti [15] introduces quantitative studies on the evolution and morphology of novels throughout history, including: visualizations for quantitative history (e.g., number of new novels per year), maps for cultural mapping (e.g., the protagonists of Parisian novels and the objects of their desire), and trees to represent evolutionary theory (e.g., evaluating the presence of clues in the early stages of British detective fiction).

The direction of his work was continued and enhanced in the Stanford Literary Lab [1, 19] via automatic tools of digital text analysis [16]. Moretti [14] also published a more comprehensive collection of essays that analyzes the morphological transformations in European novels, accompanied by research on novels' plots using network theory. His essays present quantitative information on: a) geography, as a fundamental factor in the divergence of different literary genders; b) the representation of character relations in plots, and c) statistical information on the titles, such as length or the use of adjectives and of proper names in titles.

The study by Sinykin et al. [18] was influenced by Moretti and it addresses the subjects of economics and race in American postwar novels. The study showed that women use 20% fewer economic terms than men, while African Americans use 10-15% fewer economic terms than Caucasian writers. Other studies analyze the link between book genre and writer gender in various types of writings [20], or apply cluster analysis on English novels to identify similarities in authors' writing style [7]. More in-depth studies, like the one introduced by Bode [2], analyze the evolution of the Australian novel from 1830 to the present days, the influences of other literatures, and the impact of women novelists in the national literature. The study displays statistical data, such as the number of novels by writer gender, top book publishers, places of publication, forms of publication, publisher category, genre of novels, topmost critically discussed writers etc.

The representations corresponding to the previous quantitative studies used classic visualization, such as line charts, bar charts, or manually drawn maps. Campbell et al. [4] proposed a modern representation of a collection of texts – Women Writers Online (WVO) [8] – to facilitate close and distant reading [11], and to provide easy access to general users. The WVO corpus contains more than 420 English texts written by women between 1626 and 1850, covering a wide range of topics and genres. Data representation consists of a bipartite network visualization that connects named entities to corresponding texts in which mentions can be found.

Jockers and Mimno [12] propose another modern visualization to identify how writer gender, nationality, and date of publication impact the theme of novels from the 19th century. The writers used a corpus of 3,346 novels from the 19th century covering British, Irish, and American fiction. The study is centered on identifying differences between male and female authors who write on various themes, such as: religion, war, or fashion. Moreover, the study also introduced an automated prediction of the gender of anonymous writers based on the previously generated model.

The current study is aligned with previous analyses by providing valuable insights on writers described in the DGLR through interactive visualizations. User have access in an interactive web platform to biographical information, geographical information (e.g., the countries the most important writers visited), literary entities, and publication timelines for the most important writers.

### METHOD

Our solution is a web platform that integrates several visualizations of statistical data related to Romanian writers, their writings, and other literary entities, all corresponding to letters A-O from DGLR that were currently available. The targeted writings cover domains from folklore to literary theory and expand to writings from the Republic of Moldova to writings by German, Greek, or Jewish writers on the Romanian soil. Besides writings and writers, the dictionary also includes information about editors, translators, publicists, cultural and literary movements and concepts, magazines, and cultural institutions from Romania and from Romanian diaspora, as well as anonymous writings [3]. The second edition of the dictionary is now under development; it will be available in 8 volumes, covering the information in alphabetical order. Currently, 5 volumes belonging to the second edition are already published, including letters A–O.

The data used in our visual representations was automatically extracted from DGLR and from a set of files provided separately by the Romanian Academy; these files included more detailed, granular chronological information on the life of canonical writers. A first experiment by Neagu et al. [17] was conducted on a subset of the available corpus to observe demographics of Romanian writers based on DGLR. The current work follows the same direction, processing a larger amount of data, extracting additional types of entities, and introducing novel visualizations. Additionally, the views are integrated into the ReaderBench framework and are available to the general public, as presented below.

### Indexing and Data Extraction

The corpus includes pre-print versions of DGLR together with a set of Microsoft Word documents that contain the chronology of life and literary activity of canonical writers (i.e., the most representative writers from Romanian

Literature). The information from the DGLR volumes is provided in Adobe InDesign<sup>2</sup> format, which was then converted to HTML for a standardized processing of data. The same process was applied to the Microsoft Word files which were converted to the HTML format.

The available DGLR corpus includes 2529 entities recognized as writers, from which 2433 authors were chosen for our work. The selection criterion consists of a valid year of birth that can be extracted from the description field. In addition to writers, 1186 entries were labeled as other entities (publications, associations, institutions, genres, etc.), and 1075 were included in our analysis. For the selected entries, the year of birth was found in the first line of their description using the common format “YYYY”. Disregarded entries did not have a birth year associated – for example, genres specifying only the century (“appeared in the XVIII century”). In contrast to DGLR, the corpus for the chronology of life and literary activities included only canonical writers: “Lucian Blaga”, “George Bacovia”, “Mircea Eliade”, “Constantin Dobrogeanu-Gherea”, “George Coșbuc”, “Ion Barbu”, “Tudor Arghezi”, “Mihai Eminescu”, “Emil Cioran”, and others.

Data was stored in an Elasticsearch instance, suitable for analytics purposes and fast on data retrieval in large amount of texts [9]. Two indexes were created, *index-writers* and *index-publications*, respectively, to make a separation between each category as the data stored for each entity was different. The following fields for writers were extracted from DGLR: name, year and birthplace, year, and place of death (if it is the case), professions, text biography, publishing years, list of publications and critical references. For the other literary entities, we only extracted their name and the description. Specific data preprocessing techniques were performed to extract relevant information and to structure it accordingly, as required by the graphical tools.

### Visualizations

AmCharts, a modern JavaScript library, was used to represent data. AmCharts can plot different types of views, such as: line, bar, or pie charts, as well as more complex views, such as maps, timelines, or Scalable Vector Graphics (SVG) pictorials. Besides the wide variety of views, AmCharts can render visualizations either from JSON inputs, or programmatically using the API for JavaScript or TypeScript. This was a major advantage for our application, as the standard visualizations were created using a JSON format.

The visual elements integrated in the standard charts are independent of the displayed data. As seen in Figure 1, each visualization is composed of:

- A detailed description that is displayed in the upper part of the page (1);
- A “smart” scrollbar that displays a miniature of the horizontal axis, together with two draggable bullets on each side of the scrollbar used for filtering the timeline (2);
- A cursor for better visualizing the values on the axis (3), together with tooltips available on hover for all the datapoints (4);
- A legend for each data series displayed in the chart (5);
- Labeled horizontal and vertical axis (6).

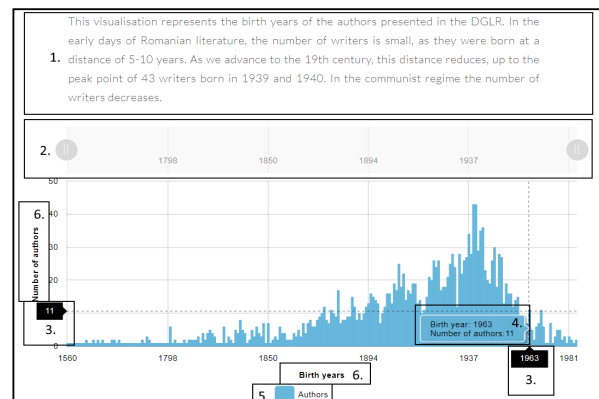


Figure 1. Anatomy of a standard chart.

Geographic maps were implemented as separate Angular components, using the AmCharts API and the geodata package<sup>3</sup>. This package contains representations of the world countries in a GeoJSON [3] format. Each map comes in two resolutions, low and high, the difference between them being the number of points used for drawing the borders. The current visualizations use the low-resolution maps, as these are faster to load. In addition, an exact border representation is not vital for our charts.

### ReaderBench Website

The ReaderBench website showcases tools for Natural Language Processing, Cohesion Network Analysis and text mining [6]. The website is developed in Angular<sup>4</sup> and is composed of numerous sections. The newly introduced visualizations were created in a separate page of the ReaderBench website, *Experiments*, centered on standalone analyses. The visualizations introduced in this paper are publicly available online, free of charge, at <http://readerbench.com/experiments/intellit>.

<sup>2</sup> <https://www.adobe.com/products/indesign.html>

<sup>3</sup> <https://www.npmjs.com/package/@amcharts/amcharts4-geodata>

<sup>4</sup> <https://angular.io/>

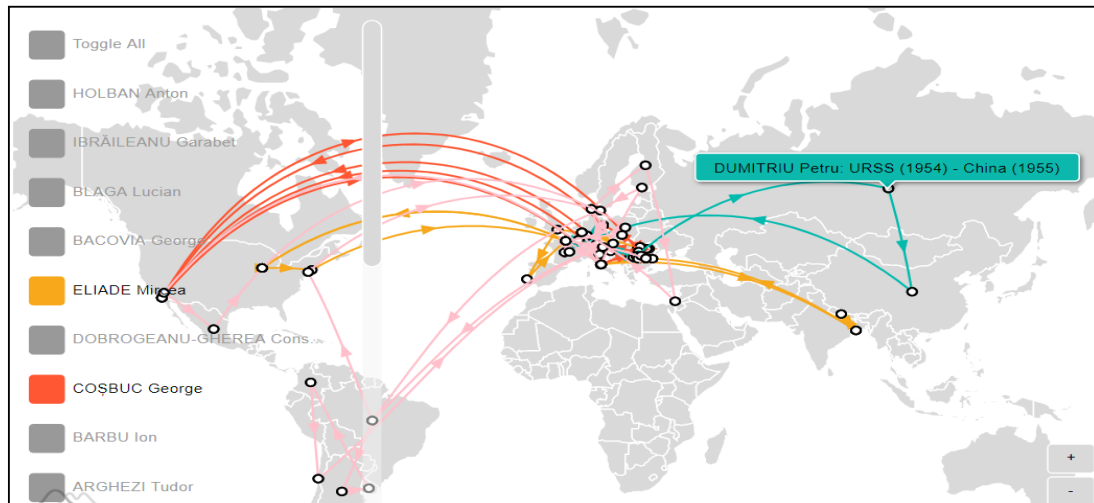


Figure 2. The travels of canonical writers.

### Experiments

The information extracted from DGLR is available to the end users as graphical visualizations divided into three categories, based on the represented data:

- *Writers throughout time* – this category contains the number of writers born each year, the birth location of the writers represented on a map, the death age of the writers, the number of publications and active writers per year;
- *Literary entities throughout time* – two views are considered, depicting the number of publications per year and the number of active publications per year;
- *Canonical writers' life* – in this category, we display the cities visited by canonical writers, their active publication period, as well as a publication timeline.

The representations can be separated into the following categories based on their type: bar-charts, maps, and miscellaneous. As a standard, the bar-charts from this experiment display on the horizontal axis the timeframe between 1515 and 2010, corresponding to the first and last recorded publication years in the DGLR. The vertical axis displays different values, depending on the selected view: the number of writers born that year, the number of publications issued, the number of active publications, or the number of active writers together with the number of published works. Another observation regarding this type of views is the existence of spikes or gaps. A line indicator depicting a 5-point moving average was applied to better highlight the trends and smooth the evolutions by filtering short-term fluctuations.

The second type of visualizations consists of two maps: the writers' birth places and the travels of canonical writers throughout the world. These types of visualizations were introduced in the initial study performed by Neagu et al. [17], but the travel map of canonical writers was enhanced, as follows (see Figure 2). The displayed paths can now be filtered by selecting and deselecting entities from the left-side legend. The "Toggle all" button removes or adds all writers from/to the map. Path directions were introduced for all travel segments, together with a tooltip displaying the start and end locations. In addition, buttons for finer zoom control were added to the bottom-right corner of the screen.

The last category of visualizations, miscellaneous, contains three visualizations. The first view is an area chart displaying the death years of writers (see Figure 3). Each year in the chart has three corresponding values: the death age the youngest and oldest authors, together with the average value between these two. Second, a miscellaneous graph considers a dumbbell plot for the publications of canonical writers [17], listing the first and last publication years. Third, we introduce an experimental timeline view, an alternative to the canonical writers' travels, available currently only for "George Coșbuc" (see Figure 6), a representative writer for Romanian literature; other authors will be added iteratively to this visualization. Each section from the timeline view is colored differently, corresponding to a time period and place where the author travelled to; the name of the place is displayed on mouse hover. The writer's publications are displayed chronographically, colored for consistency similarly to the corresponding period.

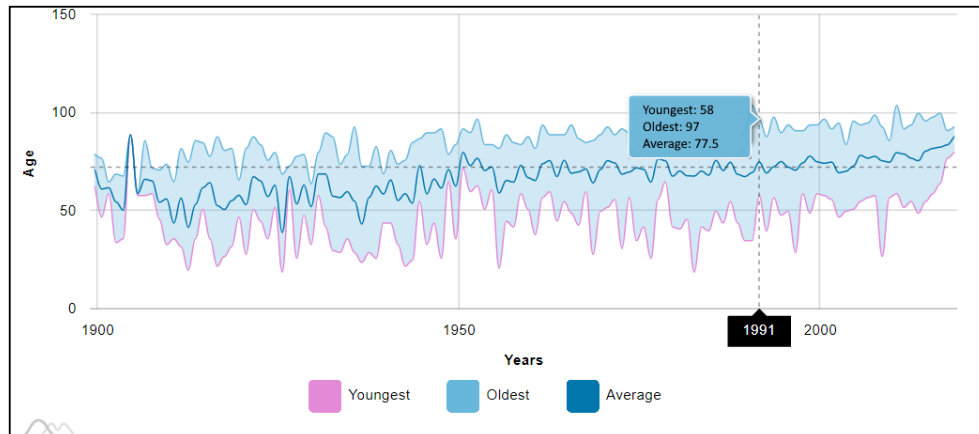


Figure 3. Age of death for writers.

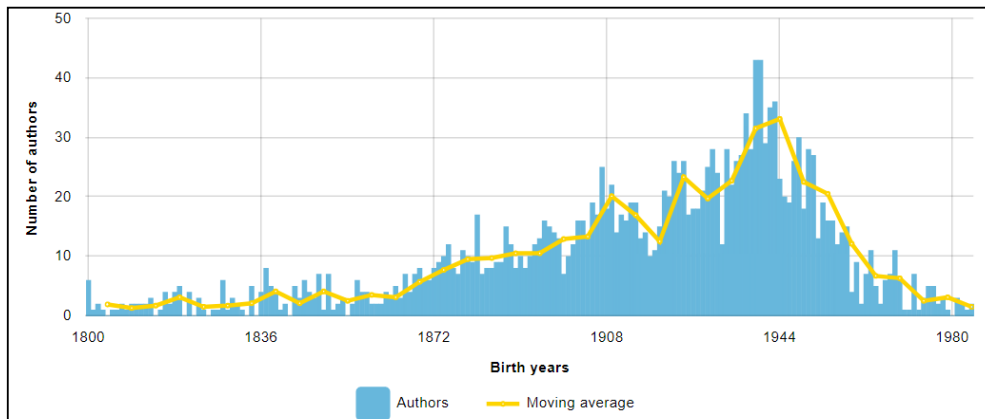


Figure 4. Writers born per year.

**DISCUSSIONS**

Our visualizations are grouped into three main categories: writers throughout time, literary entities throughout time, and the life of canonical writers. Each category has a list of visualizations presented below.

**Romanian Writers throughout Time**

The first visualization in this category aims to depict the age of the youngest and oldest writers who died every year, and the average writers’ age in the 20th and 21st centuries (see Figure 3 depicting the lowest, highest, and average value). There were many consecutive years between the 16th and 19th centuries for which there was no available data, or several years when only one writer was present. Writers before 20th century were not included in this plot. Up until 1809, only 3 years marked the death of at least 2 writers: 1711, 1715, and 1724. Data is less sparse between 1809 and 1900, but the plot would have included a lot of gaps due to the low number of writers in that period; more data was available after 1900 and until 2018.

The oldest writers in Romanian literature had 104 years (two writers), close to them one writer died at 101 years, other three at 100 years, two at 99 years, and some others at 98. At the lower end, the youngest writers who died after 1900 had only 19 years (two writers), one 20 years, and two writers were only 22. The average age of Romanian writers is 68.89 years for the full historical period included in our dataset.

The second visualization in this category (see Figure 4) highlights that the number of writers is small in the early days of Romanian literature (maximum 6), as they were born at a distance of 5–10 years; hence, data before 1800 was excluded from the analysis. The distance between the writers’ years of birth reduces starting from 1800, as we advance in the 19th century. The peak is the year 1881, when 17 writers were born; close to it are 1895 (16 writers were born) and 1887 (15 writers were born).

The observed pattern is that the Romanian literary contributions started to intensify from the middle of the 19th century. Afterwards, we observed that at least one writer was born each year in the period 1900–1980, with

the peak in 1939 and 1940 (when 43 writers were born each year), a very challenging period worldwide marking the beginning of World War II. The most flourishing period in history for the birth of Romanian writers is 1920–1951, when 17 or more writers were born each year, except 1932. Afterwards, a fall in the number of born writers was observed in the communist regime. Nevertheless, the youngest writers alive are born in 1984 (currently 36 years old); this may show that there are still future writers which are not yet well known and may fill in these gaps.

**Romanian Literary Entities throughout Time**

The first visualization in this category (see Figure 5) is related to the literary entities extracted from DGLR: literary publications, associations, and cultural institutions. Results are interesting by highlighting that the interwar period was most flourishing for the Romanian literature. Also, an important spike is shown in 1990, immediately after the communism fall, when the largest number of literary entities was encountered.

Another analysis in this category presents the literary entities active per year. The start and end years were considered the same for literary entities which had only the start year in the dictionary. There are entities which were active during certain periods of time and had missing years of activity due to wars or other internal financial problems.

**Canonical Writers Life**

A visualization in this category includes an interactive timeline of a writer’s literary activity. Figure 6 presents the timeline chart for “George Coșbuc”, a well-known Romanian writer. The timeline displays each work of the author: the work name extracted from DGLR, alongside with its corresponding publication year, together with the location where it was written.

Another analysis includes the cities visited by the canonical writers with their corresponding years (see Figure 2). This visualization shows a world map with arrows drawn between start and end cities, alongside tooltips with corresponding details.

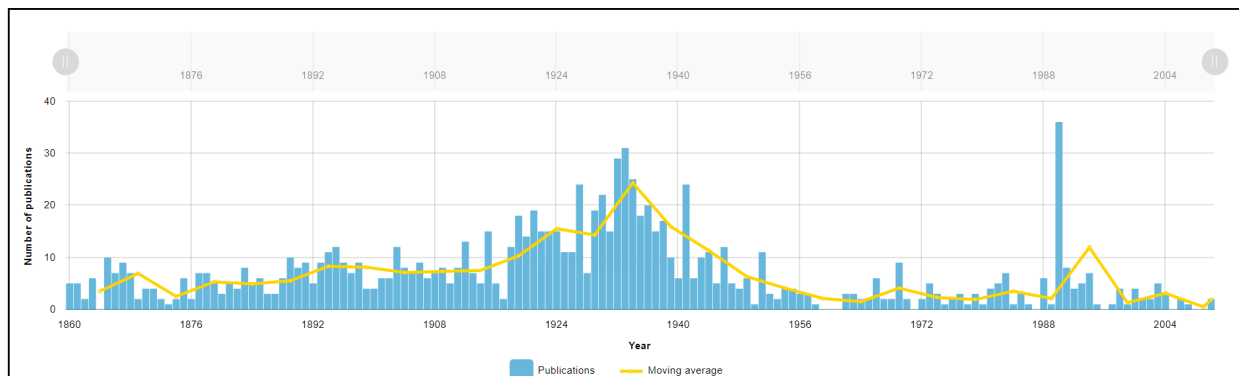


Figure 5. The number of literary entities born per year.

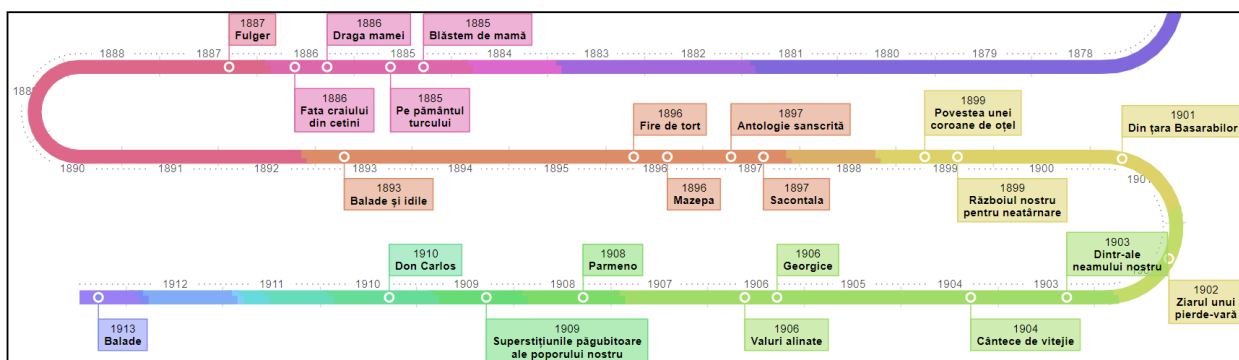


Figure 6. Timeline chart for George Coșbuc.

**USER TESTING**

A survey was conducted on 20 users, 15 males and 5 females, with ages between 25 and 45 years old. All users were asked to respond to a survey with 6 questions having ratings on a 5-point Likert Scale (1 indicates complete

disagreement and 5 complete agreement) followed by 4 free-input questions that cover their opinion on the interface and the functionalities. The first 6 questions targeted the ease of use for the more complex visualizations (i.e., canonical writers’ travels, timeline, writers’ age of death), the overall quality of the information in the interface and the usefulness of the moving average. User were asked in

their open-ended questions to provide feedback regarding the view concerning the writers' birth locations, their favorite visualization, and describe what type of information and visualizations they considered most valuable.

Two reliability statistics were calculated for the recorded answers. The Intraclass Correlation Coefficient (ICC) [13] of .624 and Cronbach's Alpha [5] of .778 denote a moderate level of agreement between the users.

The feedback received from the free-input questions was comprehensive. The first question considered the users' opinion regarding the birth location of writers represented on the map. Eighty percent of the users considered the representation interesting and useful, and 11 users requested more information on the map, such as the names of the writers and the birth years. Also, another interesting suggestion was to filter the map and display information for certain periods of time. Most users complained about the information which was too cluttered, and they experienced rendering issues. Improvements were also suggested, for example: displaying only the writers born in Romania and aggregating the rest of the writers per country. The second suggestion was to color the counties based on the number of writers born in each of them, generating a heatmap.

The second question requested users to point their favorite view and argument their choice. The results are presented in Table 1. Three users selected two views, while one user checked all visualizations as favorites. The maps and the timeline views were considered most attractive and interactive. The other charts were preferred by users who went beyond the raw information and correlated the values with historical events.

Table 1. Number of votes per type of view.

Favorite view	Number of votes
Canonical writers' travels	8
Timeline for George Coșbuc	6
Writers' birth years	3
Writers' birth places	3
Canonical writers' publications	1
Publications and active writers	1
All	1

The next two questions covered new information or visualizations that the users wanted to see in the interface. We received 22 different ideas, and we will focus on the most frequently recurring suggestions:

- Adding a visualization that represents the literary movements and the most representative corresponding writers –*five votes*;
- Extending the timeline to more writers –*four votes*;
- Introducing a tutorial for interacting with the visualizations –*three votes*;

- Adding the name of the writers/publications in views that support this feature –*three votes*;
- Depicting how events at worldwide or personal levels affected the works of the writers –*two votes*;
- Adding a view with the most important publications –*two votes*.

## CONCLUSIONS AND FUTURE WORK

The multilateral process of shifting the Romanian literature to the digital era involves literary researchers, linguists, and computer science specialists. The current paper aims to explore statistical information and web-based interactive visualizations to display data from the General Dictionary of Romanian Literature in a simple and clear way for the broad audience. The results of a survey show that most end users had a pleasant experience with our views. Future development recommendations were provided, which will be integrated in the next versions of the website.

Future work includes the integration of remaining letters from DGLR (letters P-Z), which are still under development. Moreover, the timeline view will be extended to all canonical writers. Based on the user feedback, we will address the rendering issues, add information about the writers' and other literary entities' names. Additional visualizations are envisioned, such as a heatmap for the birth places of writers, a map of the death locations of the writers, marking worldwide events on the bar-charts, as well as a short tutorial for interacting with the representations.

In terms of data sources, we plan to integrate external sources containing historical events and foreign authors, and to perform cross-correlations to observe how the social, political, and economic context influenced the Romanian writers.

## ACKNOWLEDGMENTS

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-III 54PCCDI / 2018, INTELLIT – “Prezervarea și valorificarea patrimoniului literar românesc folosind soluții digitale inteligente pentru extragerea și sistematizarea de cunoștințe” and by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

## REFERENCES

1. Allison, S., Heuser, R., Matthew, J., Moretti, F. and Witmore, M. 2011. Quantitative Formalism: an Experiment. *Stanford Literary Lab*.
2. Bode, K. 2012. Reading by numbers: Recalibrating the literary field. Anthem Press.
3. Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S. and Schaub, T. 2016. The geojson format. *Internet Engineering Task Force (IETF)*. (2016).

4. Campbell, S., Yu, Z.-Y., Connell, S. and Dunne, C. 2018. Close and Distant Reading via Named Entity Network Visualization: A Case Study of Women Writers Online. *Proceedings of the 3rd Workshop on Visualization for the Digital Humanities. VIS4DH* (2018).
5. Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*. 16, 3 (1951), 297–334.
6. Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M. and Nardy, A. 2013. ReaderBench, an environment for analyzing text complexity and reading strategies. *16th Int. Conf. on Artificial Intelligence in Education (AIED 2013)* (Memphis, USA, 2013), 379–388.
7. Eder, M. 2017. Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities*. 32, 1 (2017), 50–64.
8. Flanders, J. 2002. Learning, reading, and the problem of scale: using women writers online. *Pedagogy*. 2, 1 (2002), 49–59.
9. Gormley, C. and Tong, Z. 2015. Elasticsearch: The definitive guide: A distributed real-time search and analytics engine. O'Reilly Media, Inc.
10. Gutu-Robu, G., Sirbu, M.D., Paraschiv, I.C., Dascalu, M., Dessus, P. and Trausan-Matu, S. 2018. Liftoff - ReaderBench introduces new online functionalities. *Romanian Journal of Human - Computer Interaction*. 11, 1 (2018), 76–91.
11. Jänicke, S., Franzini, G., Cheema, M.F. and Scheuermann, G. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *EuroVis (STARs)* (2015), 83–103.
12. Jockers, M.L. and Mimno, D. 2013. Significant themes in 19th-century literature. *Poetics*. 41, 6 (2013), 750–769.
13. Koch, G.G. 1982. Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences*. S. Kotz and N.L. Johnson, eds. John Wiley & Sons. 213–217.
14. Moretti, F. 2013. *Distant reading*. Verso Books.
15. Moretti, F. 2005. Graphs, maps, trees: abstract models for a literary history. Verso.
16. Moretti, F. 2016. Literature, Measured. *Stanford Literary Lab*.
17. Neagu, L.M., Toma, I., Dascalu, M., Trausan-Matu, S., Hanganu, L. and Simion, E. 2020. A Quantitative Analysis of Romanian Writers' Demography based on the General Dictionary of Romanian Literature. *5th conference on Smart Learning Ecosystems and Regional Development (SLERD 2020)* (Bucharest, Romania, 2020).
18. Sinykin, D., So, R.J. and Young, J. 2019. Economics, Race, and the Postwar US Novel: A Quantitative Literary History. *American Literary History*. 31, 4 (2019), 775–804.
19. Stanford Literary Lab: <https://litlab.stanford.edu/pamphlets/>. Accessed: 2020-04-20.
20. Thelwall, M. 2017. Book genre and author gender: romance> paranormal-romance to autobiography> memoir. *Journal of the Association for Information Science and Technology*. 68, 5 (2017), 1212–1223.