Rupkatha Journal

On Interdisciplinary Studies in Humanities

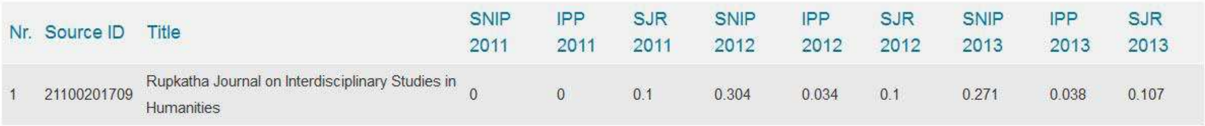
An Online Open Access Journal
ISSN 0975-2935
www.rupkatha.com

Volume VII, Number 1, 2015

Indexing and abstracting

Rupkatha Journal is an international journal recognized by a number of organizations and institutions. It is archived permanently by **www.archive-it.org** and indexed by **EBSCO**, **Elsevier**, **MLA International Directory**, **Ulrichs Web**, **DOAJ**, **Google Scholar** and other organizations and included in many university libraries.

SNIP, IPP and SJR Factors and Ranks



Nr.	Source ID	Title	SNIP 2011	IPP 2011	SJR 2011	SNIP 2012	IPP 2012	SJR 2012	SNIP 2013	IPP 2013	SJR 2013
1	21100201709	Rupkatha Journal on Interdisciplinary Studies in Humanities	0	0	0.1	0.304	0.034	0.1	0.271	0.038	0.107

Additional services and information can be found at:

About Us: www.rupkatha.com/about.php
Editorial Board: www.rupkatha.com/editorialboard.php
Archive: www.rupkatha.com/archive.php
Submission Guidelines: www.rupkatha.com/submissionguidelines.php
Call for Papers: www.rupkatha.com/callforpapers.php
Email Alerts: www.rupkatha.com/freesubscription.php
Contact Us: www.rupkatha.com/contactus.php

Digital Humanities, Big Data, and Literary Studies: Mapping European Literatures in the 21st Century

Carolina Ferrer
Université du Québec à Montréal (UQAM), Canada

Abstract

The purpose of this research is, firstly, to map the 48 national literatures of Europe, through the exploration and the analysis of the bibliographic references contained in the main literary database, the Modern Language Association International Bibliography. Secondly, the series obtained are correlated to economic and development indicators in order to determine whether and how the cultural, economic, and social fields interact with each other. From the theoretical viewpoint, this project stands at the crossroad of several concepts: the literary field defined by Pierre Bourdieu (1972, 1980, 1992), knowledge domain analysis (Hjørland& Albrechtsen 1995; Hjørland 2001; Nascimento & Marteleto 2008), scientometrics (Price 1963; Garfield 1980, 2005; Leydesdorff 1998), and the recently emerged concept of big data (Berman 2013; Boyd & Crawford 2012; Mayer-Schönberger& Cukier 2013). Methodologically, aiming at quantitatively identifying the European national literatures, we base our research on scientometrics. Initially developed by Price (1963), the purpose of scientometrics is to measure and to analyze the scientific and technological activity. In this study, we adapt scientometric indicators to the architecture and features of the Modern Language Association International Bibliography. Thus, the elaboration of bibliometric indicators (Garfield 1980, 2005; Hjorland & Albrechtsen 1995) allowed us to obtain the number of bibliographic references dedicated to the study of each of the 48 European national literatures, making it possible for us tovisualize the importance of each of these literatures and to compare them to economic and social indicators.

[Keywords: European literary field, bibliographic databases, data mining, big data, digital humanities, quantitative methods, economic indicators, social indicators]

Digital humanities and big data

In «A genealogy of digital humanities», MarijaDalbello (2011) proposesa definition of digital humanities:

> the ability to read the archive of core texts, together with their residual materiality from previous media contexts in order to produce intensive modes of engagement with particular documents, groups of texts, and the archive is brought to broader audiences. (Dalbello, 2011, p. 497)

The following year, Boyd and Crawford (2012), define Big Data «as a cultural, technological, and scholarly phenomenon that rests on the interplay of technology [...], analysis [...], mythology» (Boyd & Crawford, 2012, p. 663). The technological aspect corresponds to the capacity of extracting, storing, and putting in relation immense sets of information. Analytically, these massive amounts of information make it possible to identify patterns that allow us to obtain economic, social, and technical conclusions about the behaviour of the series. The authors consider that the belief that big datasets represent superior knowledge capable of yielding truthful, objective, and exact results is only a mythology.

Rupkatha Journal on Interdisciplinary Studies in Humanities (ISSN 0975-2935), Vol. VII, No. 1, 2015.
Ed. Tirtha Prasad Mukhopadhyay &Tarun Tapas Mukherjee
URL of the Issue: http://rupkatha.com/v7n1.php
URL of the article: http://rupkatha.com/V7/n1/01_Digital_Humanities_Big_Data.pdf
Kolkata, India. Copyrighted material. www.rupkatha.com

In 2013, Cukier and Mayer-Schoenberger (2013) establish that the phenomenon of massive information implies a change in the way we consider data. Firstly, we can no longer consider a sample of data, since huge amounts of data are available. However, this considerable amount of information implies a certain uncleanness of information. Thus, this change means, secondly, that we have to accept the existence of some inexact data, an amount that is meaningless given the quantity of information available. Finally, frequently, this data does not allow us to know the causes of the phenomena considered, allowing us only to correlate the series. Thus, there is a displacement from the determination of the causes of the events observed to their descriptions: instead of explaining the past, the correlations are used to predict the future. Moreover, as Berman (2013) points out: «Big Data provides quantitative methods to describe relationships, but these descriptions must be transformed into experimentally verified explanations» (p. 226). In this analysis, we will base our explanation on Bourdieu's study of the behaviour of the literary field (1992).

Once the definitions of digital humanities and big data established, we should examine if there is a relation between these concepts. Thus, we have extracted from the ISI Web of Knowledge the references that correspond to these concepts.

Figure 1 represents both series obtained from ISI Web of Knowledge by using the keyword technique (Callon &Penan, 1983) with the expressions "digital humanities" and "big data" in the title, for all disciplines. We observe that both series begin in the 2000s, with the exception of one publication about big data published in 1974. However, the publications about big data show a very significant growth since 2012.

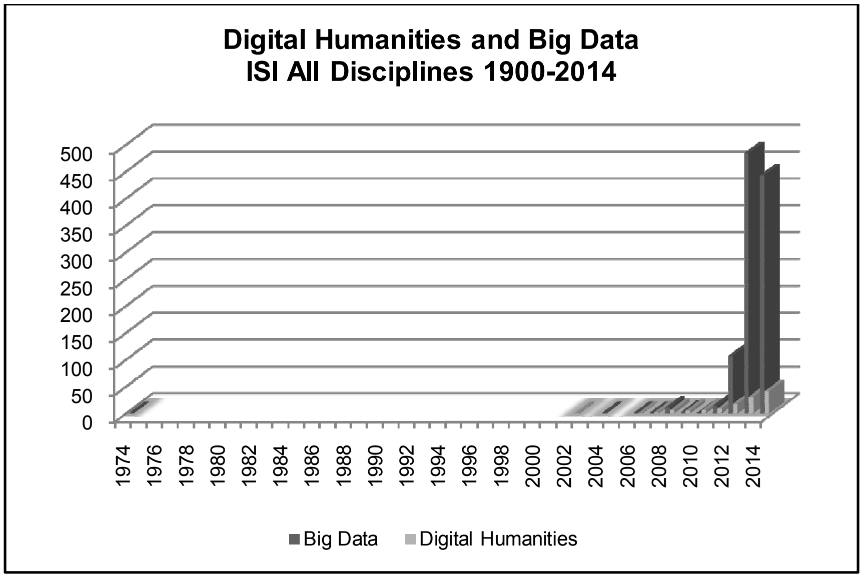


Figure 1. Digital Humanities and Big Data - All Disciplines

If we only consider the social sciences, the humanities, and the arts from the ISI Web of Knowledge database, Figure 2, we can see that the number of documents with these terms in the title is significantly smaller. These series also begin around 2000. However, the digital humanities series begins to grow in 2008. Thus, the gap between the publications about both concepts is smaller.

Moreover, according to the ISI Web of knowledge database, those documents that refer to the digital humanities belong essentially to literature (39%), library science (22%),and linguistics (18%). In contradistinction, the publications about big data correspond to business economics (25%) and to library sciences(21%). Finally, there is only one text that belongs to both series: a paper that compares quantitative and qualitative methods.

Thus, we could argue that, although both concepts share the use of computing to analyze phenomena belonging to the social sciences, the humanities and the arts, they constitute different specialties.

In spite of these observations, since 2008, we have been working on research projects that belong, to a certain extent, to both of the abovementioned tendencies. As a matter of fact, on the one hand, our research makes use of computing in order to study the behaviour of social sciences, humanities, and arts issues, particularly, literature. On the other hand, they are based on the extraction and analysis of vast amounts of data. Thus, our research considers texts as a starting point, but uses quantitative methods in order to analyze the data.

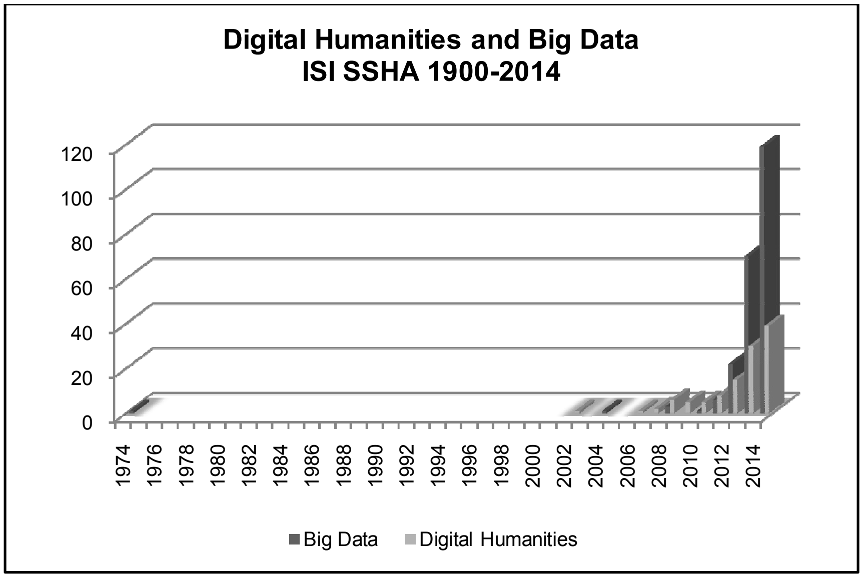


Figure 2. Digital Humanities and Big Data - Social Sciences, Humanities, and Arts

Quantitative methods and the literary field

In *Distant Reading*, Franco Moretti (2013) explains how he came to introduce quantitative methods to the analysis of novels:

> Evolution, geography, maps, series, diagrams … One step lead to the next; one step asked for the next. And one day I realized that the study of morphological evolution had itself morphed into the analysis of quantitative data. (p. 179)

Although the techniques that we use are different from those deployed by Moretti, we consider that our research has also become an analysis of quantitative data, a strange approach in literary studies. Moretti focuses on the analysis of novels, plots, titles, whereas, we study the critical bibliography about literary works, in order to analyze the field.

Again, we meet Moretti's aim at studying Europe. As he establishes:

> Where does the European novel begin? Who knows, who cares? But when it managed to survive and to grow, this is relevant, and this we know: in Europe. In the European archipelago: a space discontinuous enough to allow the simultaneous exploration of widely different paths. (p. 18)

Once more, we differ from him, since, in this paper, our purpose is to map the 48 national literatures of Europe, through the exploration and the analysis of the bibliographic references contained in the main literary database, the Modern Language Association International Bibliography, from now on MLAIB.Then, the series obtained are correlated to economic and development indicators in order to determine whether and how the cultural, economic, and social fields interact with each other. Thus, we believe that our methods and aims fairly complement Moretti's.

Theory and methodology

From the theoretical viewpoint, this project stands at the crossroad of several notions: the concept of literary field defined by Pierre Bourdieu (1972, 1980, 1992), knowledge domain analysis (Albrechtsen 1997; Hjorland & Albrechtsen 1995; Hjorland 2001; Nascimento & Marteleto 2008), scientometrics (Price 1963; Garfield 1980, 2005; Leydesdorff 1998), and big data (Boyd & Crawford 2012; Cukier & Mayer-Schoenberger 2013; Mayer-Schönberger& Cukier 2013).

According to Pierre Bourdieu (1972, 1980, 1992), society can be defined as the intertwining of fields: economic, political, religious, cultural, etc. Each field is organized according to its own logic that corresponds to the issues that characterize it. Thus, within a field, the interactions between the agents are structured according to their resources: economic, cultural, social or symbolic capital.

In their research about the disciplinary field of architecture, Nascimento and Marteleto (2008) reinforced the relations between the concept of knowledge domain analysis –developed by Hjørland and Albrechtsen (1995), Albrechtsen (1997), and Hjørland (2001)– and Bourdieu's field concept. Particularly, the authors state that:

> it becomes possible to understand how and why informational practice (as social practice) is constituted within a domain of knowledge, and, above all, interpret the historical, cultural, and social dimensions that influence the construction of information. (Nascimento & Marteleto, 2008, p. 402)

Methodologically, aiming at identifying the European national literatures in quantitative terms, we base our research on scientometrics. Initially developed by Price (1963), the purpose of scientometrics is to measure and to analyze the scientific and technological activity. Its development is due to the foundation of the Institute for Scientific Information by Eugene Garfield, nowadays internationally renowned as Thomson ISI. In this study, we adapt scientometric indicators to the architecture and features of the abovementioned MLAIB. This electronic bibliography, the most important one in literary studies, contains over 2,200,000 references and includes approximately 4,400 journals. Besides the articles, the MLAIB database includes references to books, book chapters and theses (Fitz-Enz, 2008). In terms of chronology, it covers the literary critique from 1851 to the present.

Through the techniques of data mining (Han et al., 2012; Witten et al., 2011), and keywords (Callon &Penan, 1993), we initially obtain a sample of the critical bibliography about each European national literature. Then, we extract the literary corpus as well as a list of the main writers for each country. This allows us tovisualize the importance of each of the48 European national literatures and to obtain a map of the continental literary field.

Moreover, in order to identify the tendencies of these indicators, we correlate them to economic and development indicators: the Gross National Income and the Human Development Index elaborated by the United Nations. Our purpose, at this stage, is to observe whether and how the different fields –cultural, economic, and social– interact with each other.

As abovementioned, Boyd and Crawford (2012) established that big data is «a cultural, technological, and scholarly phenomenon» (p. 663). Given the actual computational capacity of retrieving and storing massive volumes of data, it is possible to identify the existence of patterns and to correlate different series. In this sense, the elaboration of bibliometric indicators for the 48 European national literatures is an unprecedented compilation of information about the literary field.We aim at showing that this field does not stand alone, but is intertwined with other fields as Bourdieu points out.

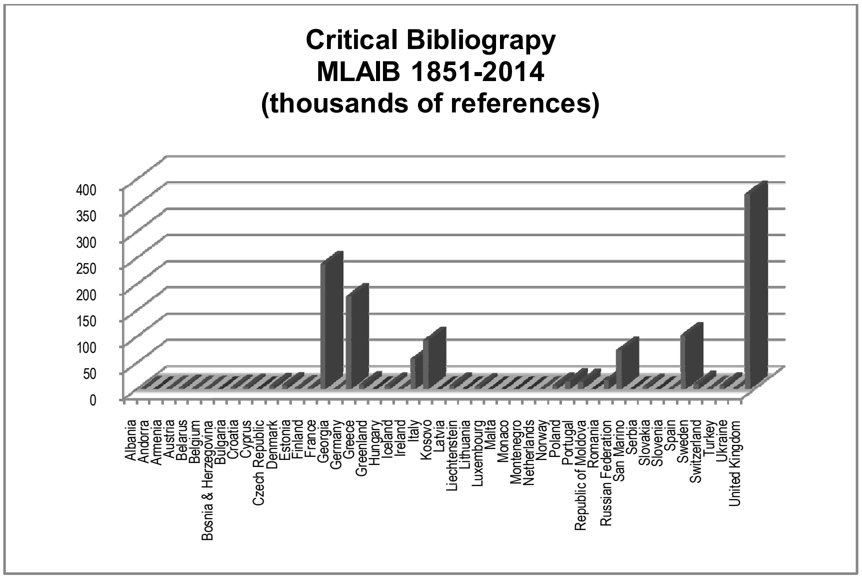


Figure 3. Critical Bibliography

The European literary field

The total set obtained for European literature contains 1,166,392 references, covering a period of 163 years, from1851 until 2014. Figure 3 represents the number of references for each of the 48 national literatures, in thousands of references. As we can observe, 7 countries clearly stand out: United Kingdom, France, Germany, Spain, Italy, Russia, and Ireland. Each of these national literatures has over 50,000 cumulated critical literary references.

Figure 4 shows the number of authors that have been the object of at least 100 publications. United Kingdom, France, Germany, and Italy have more than 100 writers in this situation.

In terms of the number of works that have 100 or more references, Figure 5, only 2 countries pass the bar of 100 titles: United Kingdom and France.

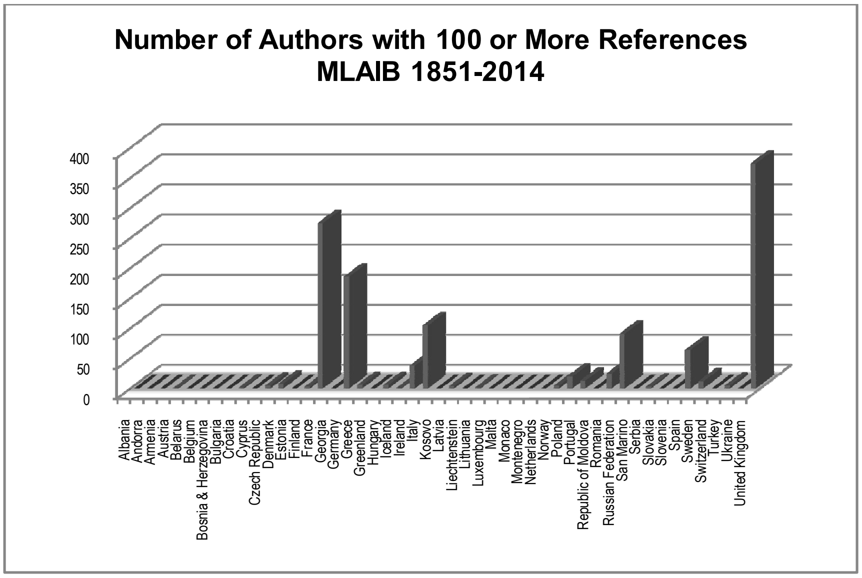


Figure 4. Number of Authors with 100 or More References

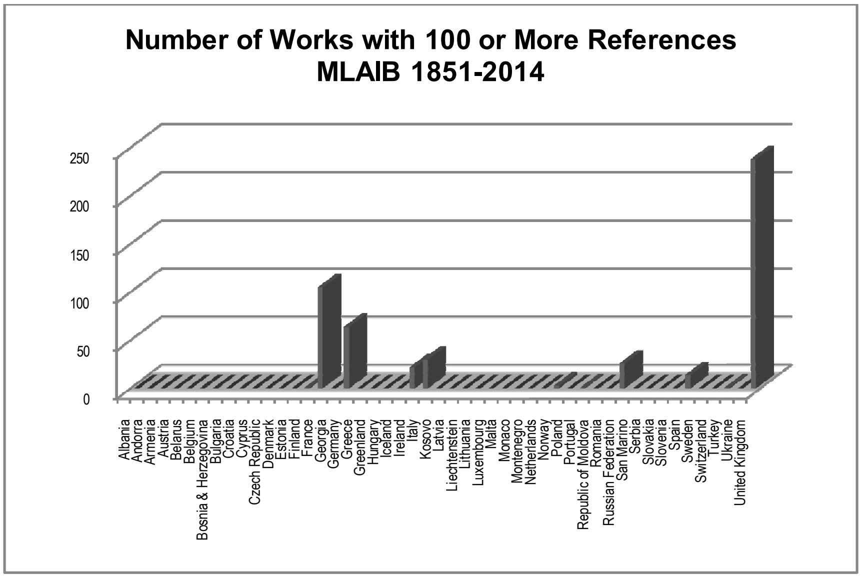


Figure 5. Number of Works with 100 or More References

Now, if we consider Table 1, we observe that only 17 countries individually represent more than 0.5% of European literature. In fact, the first 5 countries –United Kingdom, France, Germany, Spain, and Italy– cumulate 78% of the critical bibliography about European literature. Altogether, those countries with less than 0.5% of the references represent less than 3.4% of the total bibliography. This data shows that the critical bibliography about this literary field has evolved very differently across the continent. On the one hand, there are some extremely developed critical national systems, such as the abovementioned top 7 countries. On the other hand, 31 countries cumulate less than 7,000 critical references, showing an embryonic state of either the literature or the critical apparatus about it. Between these two extremes, we find 10 considerably developed national literatures, whose number of references ranges between 7,000 and 17,000 publications.

Table 1. National Literatures with 0.5% of the Continental Critical Bibliography

Rank	Country	References	Authors	Works >100	% References	Cum %
1	United Kingdom	372,280	374	238	29.5%	30%
2	France	238,715	275	105	18.9%	48%
3	Germany	177,504	186	64	14.1%	63%
4	Spain	103,035	64	14	8.2%	71%
5	Italy	93,586	105	30	7.4%	78%
6	Russian Federation	74,545	91	26	5.9%	84%
7	Ireland	57,677	38	22	4.6%	89%
8	Romania	17,022	25	0	1.3%	90%
9	Poland	14,766	20	0	1.2%	91%
10	Portugal	13,872	13	1	1.1%	92%
11	Sweden	10,314	12	0	0.8%	93%
12	Greece	9,062	6	0	0.7%	94%
13	Denmark	7,900	8	0	0.6%	94%
14	Turkey	7,402	5	1	0.6%	95%
15	Latvia	7,228	4	0	0.6%	96%
16	Norway	7,184	5	3	0.6%	96%
17	Hungary	7,043	6	0	0.6%	97%
	Countries <0.5%	42,450	17	1	3.4%	100%

Cultural, economic, and social indicators

If we now turn to socio-economic indexes, there are several indicators calculated by the United Nations (United Nations Statistics Division and United Nations Development Programme)that we may want to consider. Firstly, we can observe, Figure 6, that 9 countries have populations, in millions of inhabitants, well above the European average of 17 million people. In fact, 3 countries clearly lead in terms of population: Russia, Germany and Turkey.

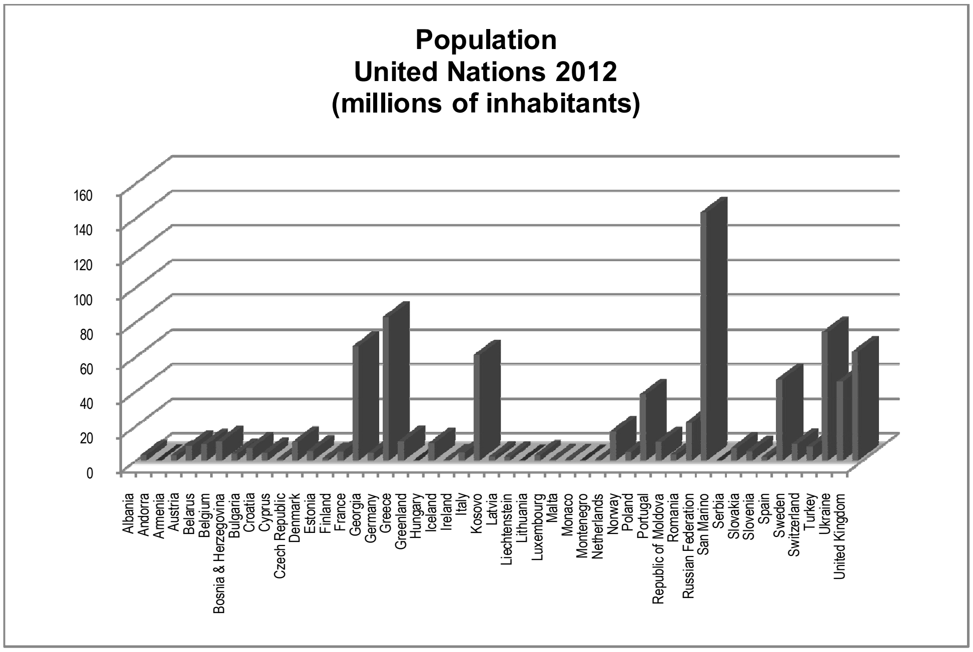


Figure 6. Population

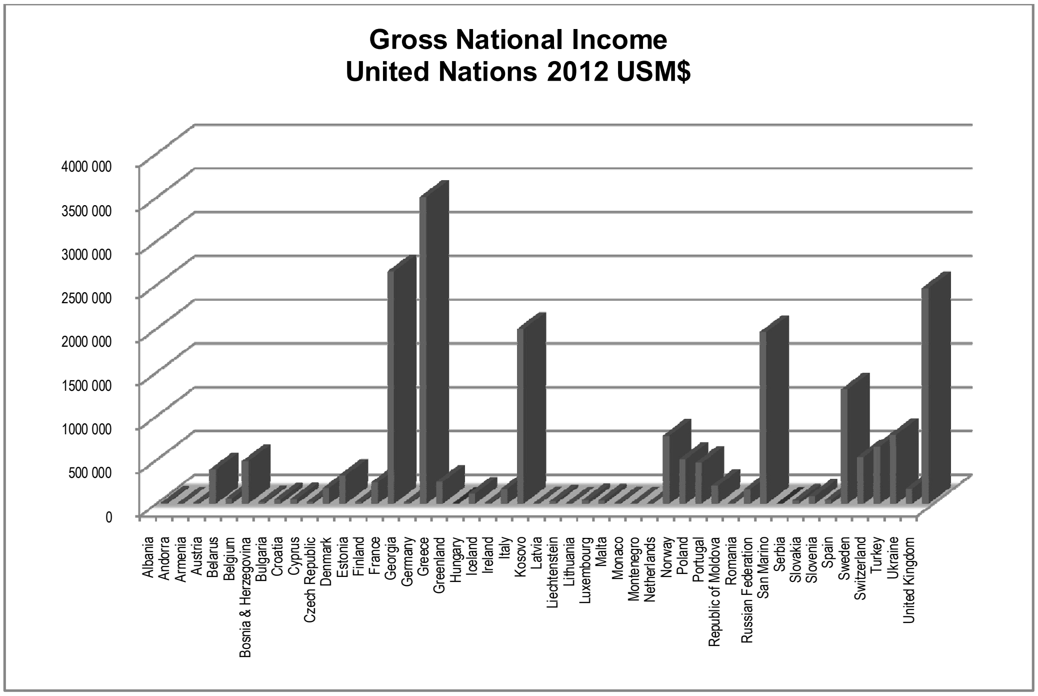


Figure 7. Gross National Income

In terms of Gross National Income, expressed in millions of 2012 US$, Figure 7, Germany, France, United Kingdom, Italy, and Russia stand out as the leaders. However, if we consider the Gross National Income per capita, the cartography is very different. In this case, Liechtenstein, Norway, and Luxembourg emerge as the countries with the highest per capita income.

Finally, Figure 8 represents the values of the Human Development Index. This indicator ranges from 0 to 1 and is a: «way of measuring development by combining indicators of life expectancy, educational attainment and income» (UNDP).Norway has the highest HDI, 0.955, whereas the Republic of Moldova holds the lowest one, 0.660. This index is not calculated for every country; thus, there is no indicator for Greenland, Kosovo, Monaco, and San Marino.

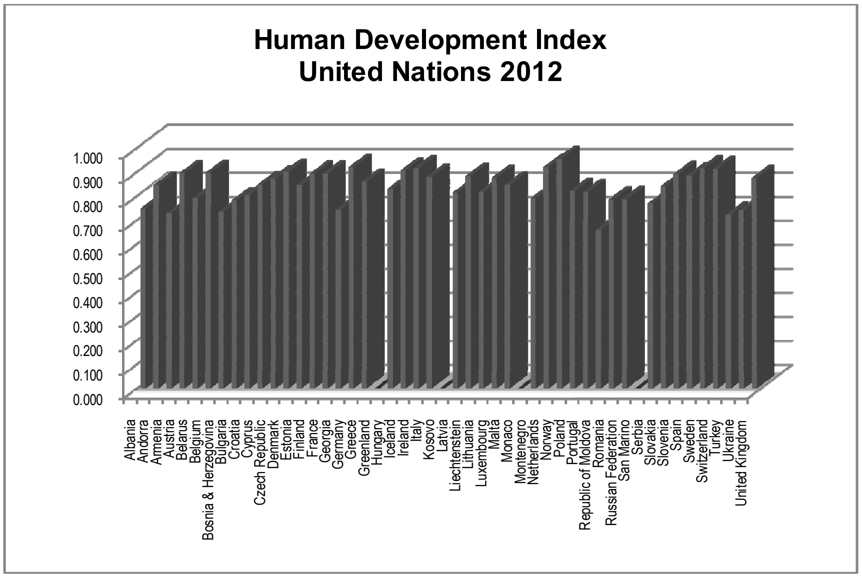


Figure 8. Human Development Index

Indicators and Literature

All along, we have been collecting data in order to answer basically one question: is there a relation between a country's cultural, economic, and social development and the importance of its national literature, measured through the critical bibliography?

In order to answer it, we calculated several correlation coefficients. Table 2 shows the results of these correlations. As we can observe, there is an important positive correlation between the Gross National Income and the number of authors with 100 references or more, 0.85, as well as between the Gross National Income and the number of references, 0.84. Also, there is a light positive correlation between the Gross National Income and the number of works with 100 references or more, 0.71, as well as between the population and the number of authors with 100 references or more, 0.62, and between the population and the number of references, 0.60. All the other coefficients are inferior to 0.50, thus, showing no relation between the indicators.

Table 2. Literary Field and Social and Economic Development

Correlation Coefficient	References	Authors>100 references	Works >100 references
Population	0.60	0.62	0.48
GNI	0.84	0.85	0.71
GNI per capita	0.14	0.14	0.14
HDI	0.24	0.23	0.20

We could then say that the level of evolution of the critical bibliography about European literatures is closely and positively related to the global level of income of these countries.

Concluding remarks

Through the exploration and analysis of the MLAIB references, we have been able to map the profile of the 48 European national literatures, in terms of the number of references, and the most studied writers and works. We have seen that the relative importance of the countries varies enormously, since some nations occupy a prominent place in the continental literary field, whereas others are at an embryonic state. Almost the same situation is observed in terms of the Gross National Income: a few countries concentrate a high percentage of the continental yearly income. Again, we observe very dissimilar population indicators. However, if we consider the Human Development Index, the spread is not so large, ranging from 0.660 to 0.955.

The most interesting results arise from the calculation of the correlation coefficients between literary and socio-economic indicators. Actually, we were able to determine that there is an important correlation between the level of development of the critical bibliography about national literatures and the Gross National Income of the European countries. Thus, we have put to a test Bourdieu's assertion about the intertwining of the social, economic and cultural fields. Essentially, we have confirmed that, in the case of the European literary field, there is a direct and positive relation between the global yearly income of a country and the level of development of the critical activity about its literature.

Finally, we believe that this study is a demonstration of the relevance of introducing quantitative methods in literary studies. Obviously, this new approach is a direct result of the recent convergence of informatics and the humanities, usually referred to as digital humanities, and the availability of big data, another recently developed concept.

References

Albrechtsen, H. (1997). Knowledge organization in the humanities.*Knowledge Organization24* (2), 61-63.

Berman, J. J. (2013).*Big data. Preparing, sharing, and analyzing complex information*. Waltham: Morgan Kaufmann.

Bourdieu, P. (1972).*Esquissed'unethéorie de la pratique, précédé de troisétudesd'ethnologiekabyle*. Genève: Droz.

Bourdieu, P. (1980).*Le senspratique*. Paris: Éditions de Minuit.

Bourdieu, P. (1992).*Les règles de l'art.Genèseet structure du champ littéraire*. Paris: Seuil.

Boyd, D., & Crawford, K. (2012). Critical questions for big data provocations for a cultural, technological, and scholarly phenomenon.*Information Communication &Society* 15.(5), 662-679.

Callon, M., &Penan, H. (1993).*La Scientométrie*. *Que Sais-Je?* 2727, Paris: Presses Universitaires de France.

Cukier, K., & Mayer-Schoenberger, V. (2013).The rise of big data. How it's changing the way we think about the world.*Foreign Affairs* 92 (3), 28-40.

Dalbello, M. (2011). A genealogy of digital humanities.*Journal of Documentation* 67 (3), 480-506.

Fitz-Enz, D. (2008).*MLA international bibliography database guide*. CSA Illumina,www.csa.com.

Garfield, E. (1980).Is information-retrieval in the arts and humanities inherently different from that in science? – Effect that ISIS-citation-index-for-the-arts-and-humanities is expected to have on future scholarship.*Library Quarterly*, 50 (1), 40-57.

Garfield, E. (2005).Identifying core literature through citation analysis and visualization.*ALA Meeting, Committee on Research and Statistics*. Chicago.

Han, J., Kamber, M.,&Pei, J. (2012).*Data mining. Concepts and techniques*. Waltham: Morgan Kaufmann.

Hjørland, B., & Albrechtsen, H., (1995). Toward a new horizon in information-science - domain-analysis.*Journal of the American Society for Information Science* 46 (6), 400-425.

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content ... and relevance.*Journal of the American Society for Information Science and Technology* 52 (9), 774-778.

Leydesdorff, L. (1998). Theories of citation?*Scientometrics43* (1), 5-25.

Mayer-Schönberger, V., & Cukier, K. (2013).*Big data. A revolution that will transform how we live, work, and think*. Boston and New York: Houghton Mifflin Harcourt.

Modern Language Association International Bibliography.www.mla.org

Moretti, F. (2013).*Distant reading*. London and New York: Verso.

Nascimento, D. M., &Marteleto R. M. (2008). Social field, domains of knowledge and informational practice.*Journal of Documentation* 64 (3), 397-412.

Price, D. S. (1963).*Little science, big science*. New York: Columbia University Press.

Thomson Reuters Web of Knowledge.http://www.isiwebofknowledge.com/

United Nations Development Programme.*Human Development Index*, http://hdr.undp.org/en/statistics/hdi.

United Nations Statistics Division.*National Accounts Main Aggregates Database*,http://unstats.un.org.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining. Practical machine learning tools and techniques*. Waltham: Morgan Kaufmann.

Carolina Ferrer is Associate Professor at the Department of Literary Studies of the University of Quebec at Montreal (UQAM), Canada.