Culture & History Digital Journal 8(1)
June 2019, e009
ISSN 2253-797X
doi: https://doi.org/10.3989/chdj.2019.009

Women and universities in *El País* (1977-2011): A methodological proposal for use of the ITCS for historical analysis

Carlos G. Figuerola[1], Tamar Groves[2] and Francisco J. Rodríguez[3]

[1]Universidad de Salamanca
e-mail: figue@usal.es
ORCID iD: https://orcid.org/0000-0001-6799-2874
[2]Universidad de Extremadura
ORCID iD: https://orcid.org/0000-0002-0494-6085
e-mail: tamargroves@unex.es
[3]Universidad de Extremadura
e-mail: franciscorj@unex.es
ORCID iD: https://orcid.org/0000-0002-2580-7981

Submitted: 11 January 2018. Accepted: 11 November 2018

ABSTRACT: The practice of historical research in recent years has been substantially affected by the emergence of the so-called digital humanities. New computer tools have been appearing, software systems capable of processing vast quantities of information in ways that until recently were inconceivable. Text mining and social network analysis techniques are sophisticated instruments that can help render a more enriching reading of the available data and draw useful conclusions. We reflect on this in the first part of this article, and then apply these tools to a practical case: quantifying and identifying the women who appear in university-related articles in the newspaper *El País* from its founding until 2011.

KEYWORDS: Digital humanities; Transition; Women; University; Social Network analysis; Text mining.

Citation / Cómo citar este artículo: Figuerola, Carlos G.; Groves, Tamar and Rodríguez, Francisco J. (2019) "Women and universities in *El País* (1977-2011): A methodological proposal for use of the ICTs for historical analysis" *Culture & History Digital Journal*, 8 (1): e009. https://doi.org/10.3989/chdj.2019.009

RESUMEN: *Mujeres y universidad en El País (1977-2011): Una propuesta metodológica para para el uso de las TIC en el análisis histórico*.- La práctica de la investigación histórica, en los años recientes, ha sido sustancialmente afectada por la aparición de las llamadas humanidades digitales. Se han introducido nuevas herramientas informáticas, sistemas de software capaces de procesar vastas cantidades de información de formas que, hasta hace poco tiempo, eran inconcevibles. Las técnicas de minería de texto y de análisis de redes sociales constituyen instrumentos sofisticados que permiten obtener una lectura más enriquecedora de los datos disponibles y extraer conclusiones útiles. Hemos reflejado esto en la primera parte de este artículo, y a continuación hemos aplicado estas herramientas a un caso práctico: cuantificar e identificar a las mujeres que aparecen en artículos relacionados con la universidad, publicados en el periódico *El País* desde su fundación hasta el año 2011

PALABRAS CLAVE: Humanidades digitales; Transición; Mujeres; Universidad; Análisis de redes sociales; Minería de textos.

In recent decades, new technologies have profoundly modified the way knowledge is produced and disseminated. Where history is concerned, computer tools have been used since the 1970s for the elaboration of databases and the selective recording of information from various sources. In addition, complete corpora of documents have been collected for subsequent automated analysis. Both processes have evolved and improved over time. The latter type consists basically in the creation of thesauri, the identification of key words and their syntactic relationships. Such exploration allows us to classify and detect the connotations and semantic patterns that underlie any message, not always perceptible in a first reading.

Since the 1990s, the extension of the internet has produced an authentic revolution in the ways of researching and generating knowledge. For the subject at hand, it highlights the possibility of accessing museum exhibits (Villaespesa, 2014) books, newspapers, and archival collections with a single click of the mouse, reducing costs and significantly increasing the possibility of locating both primary and secondary sources. At the same time, new ways of teaching, dissemination of research, and the strengthening of links between academic communities have also been facilitated by the internet (Bresciano, 2013: 45-52).

Such innovations are part of a broader phenomenon known as digital humanities (henceforth, DH). The definitions to explain it are numerous, and at times contradictory. Some highlight the dialogue between informatics and humanities (Kirschenbaum, 2010: 1-8), while others emphasize the new possibilities it offers for searching and selecting sources. According to the guide published by Northwestern University's Center for Scholarly Communication, the concept would include research presented and/or prepared by digital media, research about digital technology and culture, and even works that are critical of the way these new applications are applied.[1] Some of DH's modalities require the application of algorithms that facilitate the search, retrieval, and analysis of information.[2] Therefore specialized software is needed, though it is usually designed for this use by people not necessarily well versed in information technology. On the other hand, if one has basic knowledge of programming or can solicit the help of a computer specialist, one can adapt or create ad hoc programs for research in DH. In fact, an important feature of this area is its interdisciplinary nature, which often leads to collaboration of researchers in humanities and social sciences with experts in computing and information technology. Such collaboration must be preceded, logically, by a rigorous process of reflection and definition.

We are talking about an emerging discipline, one whose borders are difficult to define. However, those who practice it often have certain features in common. Usually, the questions addressed by DH commences from humanistic reflection, but the answers are sought through the development and application of computer tools. After all, it may be that the essential questions that preoccupy historians have not changed all that much from the days of Thucydides and Herodotus. What is unquestionable is that today's historians do their work in a totally different context—with an infinite archive, as the internet has been designated—that offers resources and possibilities unimaginable for Clio's pioneers, but which also generates new challenges. There are still many questions in the air: What kind of history will we write based on the new available sources? (Eiroa, 2011: 29).[3] Can the millions of emails that cross cyberspace every day ever be collected and studied?[4] Will we witness an "analogical desertification," or will the migration to the digital world be progressive? (Rodríguez de las Heras, 2014) And how can one confront, and winnow down, the huge overabundance of information?

WOMEN AND THE MEDIA: BETWEEN VISIBILITY AND REALITY

Starting from those questions, though by no means claiming to elucidate them completely, the following pages will try to offer a methodological proposal for using information and communications technologies (ICTs) to examine the visibility of women in relation to the university through the newspaper *El País* in the period 1977-2011. The conservative ideology of the Franco dictatorship made the incorporation of women into public life extremely difficult. It was not until the approval of the 1978 Constitution and its recognition of the principle of equality and non-discrimination based on sex that the situation of women in Spain began, slowly, to improve.[5] It was also from that point that the presence of women in the university classrooms was normalized. Previously, the figures were very unequal: in the 1950-1951 academic year, the percentage of female students was 14.8, and at its height, in 1970-71, it barely exceeded 25%. The most significant advance occurred in the period that concerns us: in just one decade, it went from just one quarter of the university population to 44% in the 1980-1981 academic year; and in 1986-1987, for the first time, the number of women pursuing higher education in Spain surpassed the number of men, at 50.1%.[6] Notwithstanding, at the levels of university research and teaching, the situation of Spanish women has not improved significantly. In the 1995-1996 academic year, only 35% of the civil servent university teaching staff was female, and only 13.2% of the full professorships were held by women.[7] In the fields of business and public administration, the situation of women in university centers has advanced even more slowly.[8]

This article will apply some of the most common IT tools within DH to gauge the extent to which these changes were reflected in the media. In addition, we hope we can facilitate the application of ICTs by other historians faced with similar challenges and that, at the same time, the collection of names can aid those who wish to pursue gender studies.

The image of women presented by the media is at the center of an ever broader academic debate. A number of studies have analyzed the role of the media in the dissemination of cultural values harmful to women: obsession with thinness (Silverstein et al. 1986: 519-532), acceptance of sexist violence, or simply the low visibility of women on these platforms (Ross, 2006). Regarding the Spanish case, a 1987 study concluded that only 8% of proper names mentioned in the press belonged to women; in 2007, they barely reached 20%. The authors of that latter study, concentrating on four digital newspapers, highlight the gap between the growing participation of women in public life and their limited presence in the media (Mateos de Cabo, 2007).

THE AUTOMATIC TREATMENT OF THE DIGITAL PRESS

One of the advantages offered by new technologies is, of course, the ability to work with huge amounts of information, and the most logical place to obtain it is the internet. The challenge is how to automatically process such a large corpus of data, whose manual and individual management would be impossible or at least would require a lifetime, and above all, how to get the most out of that information. We are not referring to access to bibliographical references, though that is also much easier thanks to the internet, but rather, the ability to work with raw primary sources and obtain knowledge from them.

One of the paradigmatic cases is that of the digital press. The trove of news, articles, and opinions published in the newspapers is, obviously, a first-hand source for the work of historians, sociologists, and political scientists. Most newspapers have had digital editions for years, apart from the paper edition; some, unable to maintain both due to economic problems, opted for electronics; others never knew ink, having been born digital. Several have digitized their pre-digital content, making enormous quantities of information available. For example, the British newspaper *The Guardian* has archives digitized since 1791; for its part, the National Library of Holland has a digital repository of 9 million pages of newspapers and around 1.5 million scientific journals;[9] even without going back that far, the interest aroused by the creation of the periodical database of the *Biblioteca Nacional* and the Historical Press webpage of the Ministry of Culture is well known.[10]

These digitalizations sometimes have drawbacks related to the way in which they were made, especially if they were encoded internally as images and not as text, which makes the computer processing of the content difficult. There are converters capable of extracting the text from the scanned images of the pages, but if they have a non-sequential layout the result will probably be less satisfactory. This is what tends to occur with *ABC* and *La Vanguardia*, a pity since they have extensive repositories. *El País* allows access to its newspaper library (from 1977 until now) through the internet; and in this case, the text of the news is encoded as such, not as a facsimile image, making it easier to operate with ICTs. Other important newspapers also have similar newspaper libraries, although with shorter temporary coverage, such as *El Mundo*, since 2002.[11]

The automatic treatment makes sense when trying to handle large volumes of information, and it is obvious that, although the sources are accessible on the internet, there is little point in manual navigation through these newspaper archives and the downloading of the news one by one. To process all this mass of information it is necessary to solve two problems: automatic download and obtaining the text of the news. For the first task, programs known as crawlers are used, which navigate autonomously through the network, from certain pages or seeds whose addresses are furnished to us based on our interests. They can also be programmed to filter what they should download and what not.[12] There are crawlers intended for specific tasks—some not so legitimate, as is the case with the collection of personal data.

For serious work of gathering information on the web, it is usual to resort to ad hoc programming; even so, there are some valid computer tools for people who are not necessarily computer experts. Probably the best known is *Wget*, a program that can be copied and used freely. Although the most common version has commands in text mode, there are versions for Windows and Mac that are easier to manage.[13] Also common are *WebSphinx* or *Heritrix*.[14]

A separate question is how to obtain the text of news so that it can be processed later. In fact, news comes over the net in different formats, usually PDF if it is a facsimile, otherwise HTML (the language of the web pages). For PDF, there are programs for conversion to plain text; The most frequent and probably the most useful when it comes to converting large amounts of documents or news is *Pdftotext*, a free program widely used.[15] Working with news in HTML format gets more complex: the conversion is more complex, since the news does not appear by itself, but is usually surrounded on a web page by other elements such as advertising, links to other news, readers' comments, etc. Consequently, a simple converter is not enough, but rather, programs capable of intelligently navigating the internal structure of a website are required, to select the part that is of interest. The programs that perform these functions are known as HTML parsers.[16]

RECOGNITION OF HTML ENTITIES

The recognition or extraction of entities (Named Entities Recognition, NER) involves locating elements in a digital text and classifying them in predefined categories, such as names of people, organizations, companies, geographical places, etc. It is a field related to natural language processing and computational linguistics, but also to statistics (Nadeau and Sekine, 2007: 3-26). This tool is highly suitable for the type of analysis that concerns us here: it allows extracting information automatically from large digital corpora, according to the parameters that we have set. This type of automated selection, which allows us to move towards possible conclusions, would take an enormous amount of time or be impossible to do manually. NER is a relatively mature technology, so it can be executed quite efficiently and does not generate very many problems. The most refined versions, however, are not available in all languages.

Though further improvements are still pending, programs currently exist for recognition of entities in Spanish with low rates of error. The lack of perfect reliability is compensated for by the advantage of being able to automatically process thousands of texts. One such system is *OpenCalais*, a product of the reputable Thomson & Reuters group, oriented towards development of the semantic web.[17] At the moment, it is operational in English, Spanish, and French. The use of it is free, although it is necessary to register and obtain a password. It works through a web API, programming that consists of communication between programs. It has a daily processing limit of around 50,000 documents.[18]

For this study, we have processed news items downloaded from the newspaper *El País* (1977-2011) containing the word *universidad* or its derivatives, such as *universitario/a*. Our objective was to find the activity of women in the universities during this period. Applying these criteria, something more than 130,000 items appeared; they had to be processed in three batches, given the aforementioned limitations of OpenCalais. The result is a list of 601,804 mentions of people detected in those news stories, corresponding to 223,543 unique people. Of them, for this work we are only interested in women. The software determines the gender of the people by first name. It is true that there are ambiguous names (Reyes, Rosario, Maria, etc.), but some of these cases are clearly exceptional and others depend on the existence of additional first names. For example, Maria is typically a female name, but it also appears in men; the difference is that in these cases it is usually preceded by another, typically male, name (for example, José María).

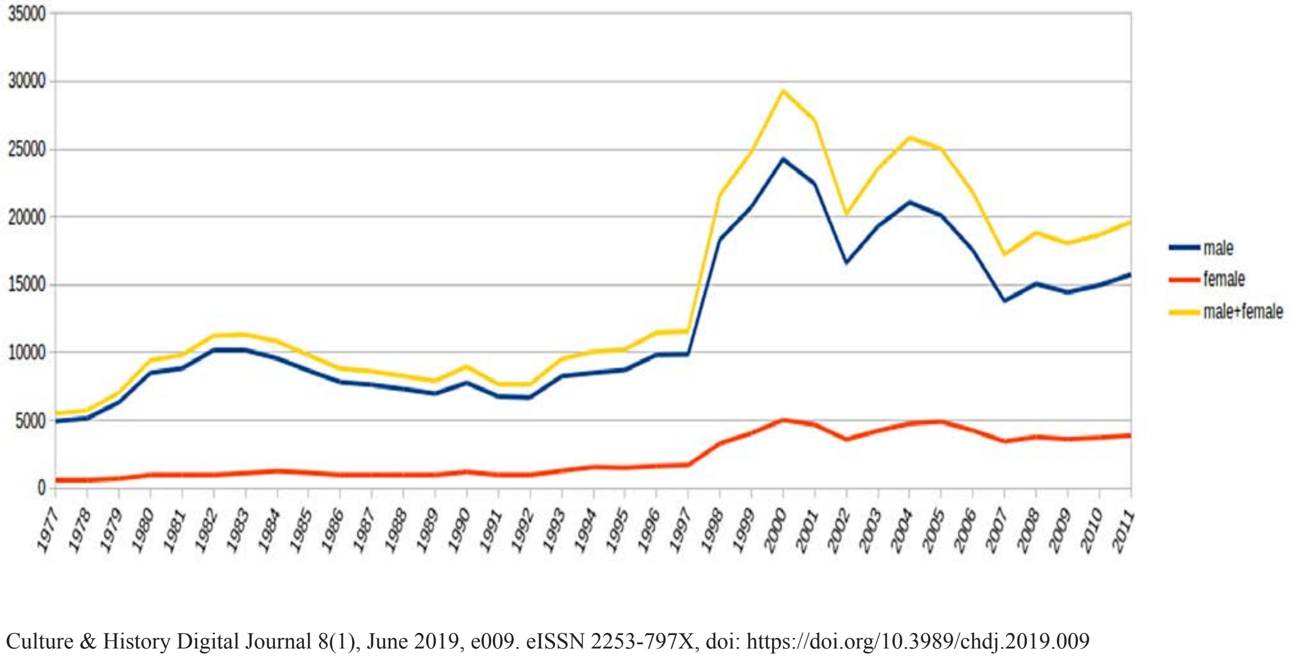
On the other hand, it could be thought that manually checking the first names of 223,543 people is an arduous task. But the first names are repeated again and again, and a simple analysis of frequencies shows us that by reviewing only 5,000 names we cover 91% of all the unique people in the entire university news corpus. From here, it is easy to calculate the number of women who appear in news related to universities each year, and observe that number's temporal evolution.

TECHNIQUES OF SOCIAL NETWORK ANALYSIS

Graph theory, a specialty of mathematics, allows us to represent abstract entities and the relationships between them and to apply different procedures and methods to such representations. This type of representation and its methods have been applied by sociologists to analyze the dynamics of social groups, but they are also used in other disciplines: from chemistry, with its atoms and links, to bibliometrics with its analysis of recurring citations. So entities and relationships can be used to model a multitude of real-world phenomena; for example, web pages and the links between them, scientific research networks, thematically divergent institutions and collaboration among them, relationships between historical figures, and relationships between words of specific texts (Hanneman, 2005).

Social network analysis technique provides us with tools to assess the importance of each of the entities represented, characterize them, determine the entire network of entities, discover indirect connections between them, and detect non-explicit communities. Thus, a network or graph consists basically of nodes or vertices that are used to represent each of the entities with which we want to work. These vertices have complementary attributes; their concrete nature depends on what we are modeling: nationality, academic institution, gender, etc. The attributes have values that depend on each vertex or concrete node. They can be connected by one or more arcs or

Graph 1. People and gender in news items related to universidad in *El País*, 1977-2011.



links, and these arcs are used to represent some kind of relationship between those nodes. We might, for example, want to model a network of researchers. Two co-authors of a scientific work, each represented by a node or vertex, could be connected by means of an arc if we were interested in analyzing collaborative networks in scientific research. The arc might or might not indicate a direction; we speak, thus, of graphs or networks being directed or non-directed. For example, in a network of citations, in which the nodes were authors and the arcs modeled or represented the events of quoting another author, such arcs would be directed: author A quotes author B, not the other way around. (Naturally, this does not prevent B from also citing A, in which case we would have another arc, but in the opposite direction.)

However, there are cases in which modeled relationships have no definite direction. Arcs or links can also have attributes; the most important, and optional, attribute is the weight of that bow; it somehow comes to represent the intensity of the relationship represented by that arc. For example, in a network of co-authorships, we could take into account the times that two authors collaborate, that is, the number of works that two people share as authors. It is evident that writing one or two articles with another person is not the same as collaborating regularly. The number of shared articles or co-occurrences would be the weight of that arc or link between such authors. Sometimes, when working with very large networks, it is common to prune the arcs or links, keeping only those arcs whose weight exceeds a certain threshold. There are two important reasons to do that: many times the excess of links entangles the final result and prevents seeing what is important in a network; also, from a practical point of view, computer programs often have difficulties with processing a large number of links.

These networks and graphs have an array of properties and characteristics that can be measured or , and from these measurements useful conclusions can be drawn. Some of these measures allow us to evaluate certain characteristics of individual nodes, while others refer to properties of the network as a whole, or parts of it. For example, certain measures tell us about the prestige or visibility of each node, its intermediation or bridge position between some parts of the network, or its greater or lesser proximity (connection) to the rest of the nodes. When working with networks of many nodes, measures of this type can allow us to distinguish sequences or factors imperceptible to the naked eye; for example, in personal networks, they help us identify members who exercise leadership but not formally, and people closely but only informally linked to known leaders. They also allow us to follow the evolution of the influence of people or groups.

Other measures are designed to evaluate the global characteristics of a network (or parts of it), such as whether it is highly interconnected or not, whether it is hierarchical or not, etc. An important part is that which allows us to detect communities of nodes, that is, groups of nodes strongly connected to each other with a weaker connection to those outside their group or community.

There are several programs to analyze social networks. Some are focused on very specific uses or aspects; others are more general and try to cover the most frequent data exploration techniques. The examples that we shall now discuss are not exhaustive or complete. We will describe several that, from personal experience or documentation in the specialized literature, we consider most useful.

Pajek

Pajek is a social network visualization program, but it also allows for calculating a wide range of measurements and coefficients, both at the node level and with a complete graph or network. It only works on Windows and, due to its internal technology, is capable of handling large networks, precisely the area where many other programs show more weakness. Its interface, although graphic and with windows, is not quite intuitive, perhaps due to the complexity of the tasks it is capable of performing. Fortunately, it is well documented; the classic book *Exploratory Social Network Analysis with Pajek* is the official program manual, while *Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution* deals more directly with its use and application. Although it is not, strictly speaking, distributed as free software, it can be downloaded and used without restrictions.[19]

Gephi

Gephi is a more modern software, with a clearer interface in terms of its comprehension. It is also a generalist program, in the sense that it tries to cover almost all known techniques of social network analysis. It produces graphics of higher quality than Pajek, and it is also possible to interact with those graphics. It is capable of importing and exporting data in a multitude of formats, including the importing of spreadsheets. One of its strengths is the ability to analyze diachronic networks, even producing animations with the evolution of these networks. Gephi can work with large networks, but it will need more memory than Pajek. On the other hand, it works on Windows, Linux, and Mac; and yes, it is indeed free software. Its source code is available, and there are currently numerous complementary tools developed by its large community of users.[20]

Ucinet

Ucinet is an older program: its first versions date from the 1980s. After numerous vicissitudes and versions, it landed in Windows in 2002. Although its interface is much less flashy than that of other programs, it is one of the most complete, especially suitable for working with large networks. It has been used by numerous researchers in various fields for many years, making it easy to find documentation on its use. Unlike some others, Ucinet is not free software, although its price for academic uses is affordable. However, there are few extensions or supplements developed by third parties.[21]

Vennmaker

The program *Vennmaker* is much less well known, but we include it here because it has been used in topics related to historiographic research. This software is not free.[22]

FROM THEORY TO PRACTICE

From news items in *El País* (1977-2011) dealing with the university, and thanks to Recognition of Entities software, we can identify the women who appear in these news items. We can model or represent information through social networks, with each woman considered as a node in that network. In the same way, we could think that if two women are mentioned in the same news, there is some kind of relationship or link between them: they participated in the same event or activity, they work in the same field, etc. Following that same approach, the strength or weight of that relationship will be proportional to the number of times that both women appear together: the more events and situations these women take part in together, the more weight the link will have. It could be objected—not without reason—that sometimes that co-occurrence in the same news is the result of chance, or due to circumstances that are merely conjunctural. However, when we work with such large amounts of data, the casual coincidences become irrelevant: two women could appear together in one or two stories by chance, but if those women appear together in a thousand news items, as is shown in Table 1, it cannot be a fortuitous phenomenon.

Let us now look at some of the details in the data gathered. We detected the presence of 80,253 women in a total of 133,738 news items related to Spain's universities in the pages of *El País*. In those figures we have identified 35,996 different women. But only 90 women appear in 50 or more items, while 27,190 appear in only one.

If we construct our network only with women who, in the whole period, appear in 5 or more news items, we obtain a graph of 2,278 nodes (women) and 11,172 edges. Table 2 shows women with a higher degree, a simple measure of prestige or visibility in a network, while Table 3 shows the women with the highest Page Rank, a far more sophisticated measure of the importance or influence of a node in the network.

Table 1. The women who appear in the most news items, searched by the keyword *universidad* (*El País*, 1977-2011)[23]

Name	No. of news items
Esperanza Aguirre	798
María Teresa de la Vega	655
Pilar del Castillo	602
Sofía	530
Margaret Thatcher	373
María L. Rodríguez Tapia	358
Angela Merkel	210
María Zambrano	190
Carmen Calvo	185
Carmen Martín Gaite	178
Rita Barberá	173
Elena Salgado	169
Carmen Alborch	165
Mercedes Cabrera	161
María J. San Segundo	155

Table 2. Women with higher degrees in news items about *universidad* (*El País*, 1977-2011)[24]

Esperanza Aguirre
María Teresa de la Vega
Sofía
Carmen Martín Gaite
Almudena Grandes
Carmen Alborch
Carmen Calvo
Pilar del Castillo
Ana María Matute
Ana Belén
Ana Botella
Carmen Iglesias
Margaret Thatcher

Table 3. Women with higher Page Rank in news items about *universidad* (*El País*, 1977-2011)[25]

Esperanza Aguirre
Sofía
María Teresa de la Vega
Pilar del Castillo
Carmen Martín Gaite
Carmen Alborch
Almudena Grandes
Carmen Calvo
María L. Rodríg. Tapia
Ana María Matute
Margaret Thatcher
Carmen Iglesias
Ana Botella

TABLE 4. Principal communities of the first order, and members with higher degrees

Community A	Community B	Community C	Community D	Community E
Sofía	Carmen Martín Gaite	Amelia Valcárcel	Rita Barberá	Montserrat Caballé
Cristina Iglesias	Almudena Grandes	Hannah Arendt	Ana Noguera	Teresa Berganza
María Corral	Ana María Matute	Adela Cortina	Consuelo Ciscar	Alicia de Larrocha
Juana de Aizpuru	Carmen Iglesias	Celia Amorós	Marguerite Yourcenar	Barbara Hendricks
Ana Rossetti	Carmen Balcells	Mary Robinson	Carmen Martorell	Arantxa Sánchez Vicario
Carmen Giménez	Margarita Salas	María Emilia Casas	María Fabra	María Callas
Carmen Laffón	Rosa Montero	Elisa Pérez Vera	Marina Baixa	Ainhoa Arteta
Carmen López	Ana María Moix	Ana María Pérez del Campo	Rosa Serrano	Aretha Franklin
Carmen Garrido	Pilar Miró	Enriqueta Chicano	Alicia de Miguel	Patrice Chéreau
Isabel Ignacio	Rosa Regàs	Molly Bloom	Marcela Miró	Montserrat Torrent
Pilar López	Elvira Lindo	Martha Nussbaum	Susana Camarero	María Callas
Community F	**Community G**	**Community H**	**Community I**	**Community J**
Margaret Thatcher	Esperanza Aguirre	Ana Belén	Reyes Magos	Virginia Woolf
Cristina	María Teresa de la Vega	Nuria Espert	Ana Patricia Botín	Angela Merkel
Elena	Carmen Alborch	Concha Velasco	Amparo Moraleda	María L. Rodríg. Tapia
Diana de Gales	Carmen Calvo	Charo López	Almudena Fontecha	Susan Sontag
Cristina Fernández	Pilar del Castillo	Aitana Sánchez-Gijón	Esther Koplowitz	Marilyn Monroe
Isabel Burdiel	Ana Botella	MaríaGuerrero	Rocío de Sevilla	Ava Gardner
Isabel Ferrer	Trinidad Jiménez	Carmen_Linares	Carmen Bravo	Isabel Coixet
Carmen Sarmiento	Cristina Narbona	Rosa León	Nuria Chinchilla	Nicole Kidman
Bibiana Fernández	Cristina Alberdi	Lola Flores	Isabel García Marcos	Carla Bruni
María Kodama	Rosa Díez	Carmen Maura	Cecilia Castaño	Sarah Palin
Aurora Bernárdez	Leire Pajín	Marisa Paredes	María Dolores	Barbara Probst Solomon

More interesting is the formation of communities, that is, groups of women who tend to appear in the same news frequently. Visualization of the network using a force-directed system (that is, trying to put together the nodes that show more intensity or weight in their linkages) (Alonso, Figuerola, Medrano, 2011: 167-190) allows us to appreciate more clearly the existence of several communities (Figure 1).

If we resort to a relatively simple system of community detection, modularity (Blondel, 2008), and without deepening or going beyond a generic first level, we obtain the communities that appear in table 4. In the two columns, we have arranged the larger-sized communities detected and, for each one of them, the women with the higher degrees.

PRELIMINARY CONCLUSIONS

As we noted in the introduction, we are working in a field, that of the digital humanities, that is very much a work in progress with software in constant evolution, so it would be risky to assert any definitive conclusions. However, we can advance some provisional findings while marking possible future lines of action. In the first place, the scarce presence of women in news items related to the university is confirmed. As shown clearly in graph 1, there is a very significant gap between the number of men and women who appear in news related to the university. The situation began to improve, albeit slowly, from the end of the 1990s, but there is still a wide gap. Secondly, we observe that the women who appear in uni-

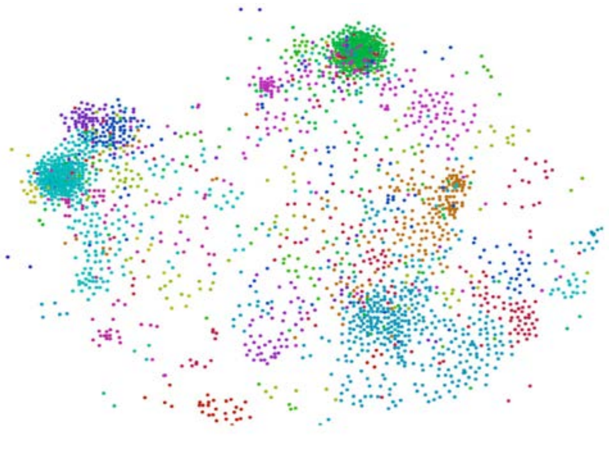


FIGURE 1. Force-directed visualization.

versity-related news come from the worlds of politics or literature: women who occupy prominent positions in public life, but not necessarily directly related to higher education. The percentage of those who actually dedicate themselves specifically to university-centered tasks is, on the contrary, very small. Our verification of that point corroborates the lack of visibility of university women in the press. On another occasion, these figures will have to be compared with those of male researchers and teachers, which will lead to some other questions to explore: In which fields is the "invisibility" of women greater, and why? And in the opposite direction: what unites women who did have greater visibility? What is their academic background? Their political affiliation? On the other hand, with respect to the location of communities, there appear groups of women apparently not related to each other (see table 4). However, as we explained, it is difficult to argue that coincidence is the result of chance when they appear together on hundreds of occasions. We will have to examine the internal logic of these communities at another time. In short, we are facing a terrain scarcely charted out by historiography that needs to go further. These pages aspire to be a marker of this new path.

Information technologies can provide an advantage in historiographical research. With the increasing availability of information on the internet, it is necessary to consider how to work with such extensive documentary corpus, with data sometimes so abundant that the trees prevent you from seeing the forest. Advances in computing allow for automatically locating, collecting, and selecting a good part of all that information and establishing the parameters that best suit what we want to investigate. Application of the software explained here requires some training. Natural language processing, text mining, and social network analysis techniques are sophisticated tools with a degree of difficulty that is sometimes high, but which can help us to read the available data more enrichingly, generate new knowledge, and draw useful conclusions The great versatility and potential of these techniques should not hide the challenges and difficulties that they also entail—basically, how do you interpret the information extracted?

Finally, teamwork has allowed us to reflect on how historical research (and academic research in general) should be understood as an interdisciplinary issue. The multidisciplinary team, however, should not be considered as a mere juxtaposition of watertight disciplines, but as a place of scientific miscegenation, where differences enrich and act as the key that allows for greater advance.

NOTES

[1] http://libguides.northwestern.edu/dh. [consulted February 9, 2015]

[2] Schreibman, Siemens and Unsworth (2004) *A Companion to Digital Humanities*. Oxford. Blackwell.

[3] Eiroa (2011) "Historia digital, historia de los medios digitales: antiguos dilemas para nuevos paradigmas". *Conexiones,* vol. 3, no. 2: 29.

[4] Google vice-president Vint Cerf warned that "Piles of digitised material–from blogs, tweets, pictures and videos, to official documents such as court rulings and emails–may be lost forever.... Humanity's first steps into the digital world could be lost to future historians". *The Guardian,* February 13, 2015, available at: http://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf?CMP=fb_gu [consulted February 13, 2014]

[5] A recent contribution on women during the Spanish transition, which also includes the use of ICTs, is Moreno Seco, Mónica (2013) "Presentación. Mujeres de la Transición" *Cuestiones de género: de la igualdad y la diferencia*, vol. 8: 5-7. (It is a monographic volume on this subject.)

[6] Flecha García (2014) "Desequilibrios de género en educación en la España Contemporánea: causas, indicadores y consecuencias". *Revista Internacional de Ciencias Sociales* vol. 33: 49-60; Sanchidrián Blanco Carmen (2008) "Estudios Universitarios y ejercicio profesional de las mujeres en el Franquismo". En Jiménez Fernández and Pérez Serrano (coords.), *Educación y género. El conocimiento invisible.* Valencia. UNED and Tirant lo Blanch.

[7] European Commission: *Science Policies in the European Union: Promoting Excellence through Mainstreaming Gender Equality*. Luxembourg. European Commission, 2000.

[8] Groves, López Santiago and Gutiérrez Palmero: "El último impulso: mujer y ciencia en las universidades de Castilla y León entre el tardofranquismo y la democracia". In Cuesta , de Prado y Rodríguez (eds.) (2015) *¿Mujeres Sabias? Mujeres universitarias en España y América Latina*. Limoges. Presses Universitaires de Limoges.

[9] Pieters and Verheul (2014) "Cultural text mining: using text mining to map the emergence of transnational reference cultures in public media repositories". Available at http://dharchive.org/paper/DH2014/Paper-757.xml [consulted December 17, 2014]. An interesting reflection on the risks and opportunities of these huge digital corpora is found in Bingham: "The digitization of newspaper archives: Opportunities and challenges for historians". *Twentieth Century British History*, vol. 21: 225-231.

[10] http://prensahistorica.mcu.es/es/estaticos/contenido.cmd?pagina=estaticos/presentacion, [consulted August 5, 2018].

[11] http://www.elmundo.es/elmundo/hemeroteca [consulted February 9, 2015].

[12] Castillo (2004) *Effective Web Crawling* (Doctoral Thesis). Available at http://chato.cl/research/crawling_thesis [consulted September 2, 2015]; Olston and Najork (2010) "Web Crawling: Invited Survey Article" *Journal of Foundations and Trends in Information Retrieval*, 4: 175-246.

[13] https://www.gnu.org/s/wget [consulted February 9, 2015].

[14] http://www.cs.cmu.edu/~rcm/websphinx. Miller, Robert and Bharat (1998) "SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers". *Computer Network and ISDN Systems*, vol. 30: 119-130. https://webarchive.jira.com/wiki/display/Heritrix. [consulted January 14, 2015].

[15] http://www.foolabs.com/xpdf/download.html. Some of its characteristics are discussed in Chen and Gey (2003) "Combining query translation and document translation in cross-language retrieval". In Peters (et al., eds.) *Comparative Evaluation of Multilingual Information Access Systems*. Berlin: Heidelberg, CLEF 2003: 108-121.

[16] A fairly illustrative description of these processes is found in Sahuguet and Azavant (1999) "Building light-weight wrappers for legacy web data-sources using W4F". *VLDB*, vol. 99: 738-741.

[17] http://www.opencalais.com. [consulted February 9, 2015]

[18] http://en.wikipedia.org/wiki/Web_API [consulted December 14, 2015].

[19] de Nooy, Mrvar and Batagelj (2010) *Exploratory Social Network Analysis with Pajek.* Cambridge. Cambridge University Press. Batagelj and Kejžar (2014) *Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern*

Searching, Visualization and Network Evolution. New York. Wiley.

20 Gephi reaches a wide clientele, so there is abundant documentation on the internet about its operation and characteristics. The most recommended book on this is *Network Graph Analysis and Visualization with Gephi*. More details are on its website: http://gephi.org. Cherven (2013) *Network Graph Analysis and Visualization with Gephi*. Birimingham. Packu Publishing.

21 We know of the existence of at least one manual for this program: Borgatti, Everett and Johnson (2013) *Analyzing Social Networks*. London. Sage Publications. The website is: https://sites.google.com/site/ucinetsoftware/home.

22 Stark and Kronenwett (2014) “Digital Humanities und Digitale Netzwerkkarten. Eine Software für die Geisteswissenschaften, Posterpräsentation auf der DHd Passau Jahrestagung Digital Humanities –methodischer Brückenschlag oder ‘feindliche Übernahme?”. Available at: http://www.dhd2014.uni-passau.de/. Düring, Bixler, Kronenwett and Stark (2015) “VennMaker for Historians: Sources, Social Networks and Software”. Available at: http://revista-redes.rediris.es/html-vol21/vol21_8e.htm [consulted February 14, 2015]. Vennmaker’s website is at: http://www.vennmaker.com/.

23 For this table, we selected women who appeared in more than 150 news ítems.

24 The individuals named in the left-hand column have higher degrees than those on the right.

25 See preceding note.

REFERENCES

Alonso, José Luis; Figuerola, Carlos and Medrano, José F. (2011) “Visualización de grafos Web”. In *Avances en Informática y Automática. Quinto Workshop*. Salamanca. Dpto. de Informática y Automática: 167-190.

Batagelj, Vladimir and Kejžar, Nataša (2014) Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. New York. Wiley, October.

Bingham, Adrian (2010) “The digitization of newspaper archives: Opportunities and challenges for historians”. *Twentieth Century British History*, vol. 21, February: 225-231.

Blondel, Vincent; Guillaume, Jean-Loup; Lambiotte, Renaud and Lefebvre, Etienne (2008) “Fast unfolding of communities in large networks”. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10, available at: http://arxiv.org/abs/0803.0476 [consulted February 18, 2015].

Borgatti, Stephen; Everett, Martin and Johnson, Jeffrey (2013) *Analyzing Social Networks*. London. Sage Publications.

Bresciano, Juan Andrés (2000) La investigación histórica y las nuevas tecnologías. Montevideo. Librería de la Facultad de Humanidades y Ciencias de la Educación.

Castillo, Carlos (2015) Effective Web Crawling (Doctoral Thesis, 2004), available at http://chato.cl/research/crawling_thesis [consulted September 2].

Chen, Aitao and Gey, Frederic (2003) “Combining query translation and document translation in cross-language retrieval”. In Peters, Charles et al., (eds.). *Comparative Evaluation of Multilingual Information Access Systems*. Berlin. Heidelberg, CLEF: 108-121.

Cherven, Ken (2013) *Network Graph Analysis and Visualization with Gephi*. Birimingham. Packu Publishing.

de Nooy, Wouter, Mrvar, Andrej and Batagelj, Vladimir (2010) *Exploratory Social Network Analysis with Pajek*. Cambridge. Cambridge University Press.

Düring, Marten; Bixler, Matthias; Kronenwett, Michael and Stark, Martin (2011) “VennMaker for Historians: Sources, Social Networks and Software”. *REDES-Revista hispana para el análisis de redes sociales*. Vol. 21, no. 8. Available at: http://revista-redes.rediris.es/html-vol21/vol21_8e.htm [consulted February 14].

Eiroa, Matilde (2011) “Historia digital, historia de los medios digitales: antiguos dilemas para nuevos paradigmas”. *Conexiones*, vol. 3, no. 2: 21-36.

European Commission (2000) *Science Policies in the European Union: Promoting Excellence Through Mainstreaming Gender Equality*. Luxembourg. European Commission.

Flecha García, Consuelo (2014) “Desequilibrios de género en educación en la España Contemporánea: causas, indicadores y consecuencias”. *Revista Internacional de Ciencias Sociales*, vol. 33, 49-60.

Groves Tamar, López Santiago M., and Gutiérrez Palmero, Mª José (2015) “El último impulso: mujer y ciencia en las universidades de Castilla y León entre el tardofranquismo y la democracia”. In Cuesta Josefina, de Prado María Luz and Rodríguez Francisco J. (eds.) *¿Mujeres Sabias? Mujeres universitarias en España y América Latina*. Limoges. Presses Universitaires de Limoges.

Hanneman, Robert and Riddle, Mark (2005) “Introduction to social network methods”. Riverside. University of California.

Kirschenbaum, Matthew (2010) “What Is Digital Humanities and What’s It Doing in English Departments?”. *ADE Bulletin* no. 150: 1-8.

Mateos de Cabo, Ruth (coord.) (2007) *La presencia de estereotipos en los medios de comunicación: análisis de la prensa digital española*. Madrid. Comunidad de Madrid.

Miller, Robert and Bharat, Krishna (2015) “SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers”. *Computer Network and ISDN Systems*, vol. 30 (1998): 119-130. http://www.cs.cmu.edu/~rcm/papers/www7/www7.html. [consulted January 14].

Moreno Seco, Mónica (2013) “Presentación. Mujeres de la Transición”. *Cuestiones de género: de la igualdad y la diferencia*, vol. 8: 5-7.

Nadeau, David and Sekine, Satoshi (2007) “A survey of named entity recognition and classification”. *Lingvisticae Investigationes*, vol. 30 (January), 3-26.

Olston, Christopher and Najork, Marc (2010) “Web Crawling. Invited survey article”. *Journal of Foundations and Trends in Information Retrieval*, 4, 175-246.

Pieters, Toine and Verheul, Jaap (2014) “Cultural text mining: using text mining to map the emergence of transnational reference cultures in public media repositories”. Available at http://dharchive.org/paper/DH2014/Paper-757.xml [consulted December 17].

Rodríguez de las Heras, Antonio: “La migración digital”. *Telos*, no. 61.

Ross, Karen, and Byerly (eds.) (2006) *Women and media: international perspectives*. Oxford. Blackwell.

Sahuguet, Arnaud and Azavant, Fabien (1999) “Building lightweight wrappers for legacy web data-sources using W4F”. *VLDB*, vol. 99, 738-741.

Sanchidrián Blanco, Carmen (2008) “Estudios Universitarios y ejercicio profesional de las mujeres en el Franquismo”. In Jiménez Fernández, Carmen y Pérez Serrano, Gloria (coords.), *Educación y género. El conocimiento invisible*, Valencia. UNED and Tirant lo Blanch.

Schreibman, Susan, Siemens, Ray and Unsworth, John (2004) *A Companion to Digital Humanities*. Oxford: Blackwell.

Silverstein, Brett, et al. (1986) The role of the mass media in promoting a thin standard of bodily attractiveness for women, *Sex Roles*, vol. 14, no. 9/10: 519-532.

Stark, Martin; Kronenwett, Michael (2014) “Digital Humanities und Digitale Netzwerkkarten. Eine Software für die Geisteswissenschaften, Posterpräsentation auf der DHd Passau Jahrestagung Digital Humanities –methodischer Brückenschlag oder ‘feindliche Übernahme’”? Available at: http://www.dhd2014.uni-passau.de/.

Villaespesa, Elena (2014) “Comunicación 2.0: Tecnologías de la Información y uso de Redes Sociales en los Museos”.