ВОПРОСЫ ОНОМАСТИКИ
PROBLEMS OF ONOMASTICS

Xulio Sousa
Instituto da Lingua Galega, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Geonomastics on the Web: Visualizing Surname Distributions in a Regional Space

Voprosy onomastiki, 2019, Vol. 16, Issue 3, pp. 212–223
DOI: 10.15826/vopr_onom.2019.16.3.041

Language of the article: English

Xulio Sousa
Instituto da Lingua Galega, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Geonomastics on the Web: Visualizing Surname Distributions in a Regional Space

Вопросы ономастики. 2019. Т. 16. № 3. С. 212–223
DOI: 10.15826/vopr_onom.2019.16.3.041

Язык статьи: английский

Ural Federal University
named after the first President of Russia B.N.Yeltsin

ИНСТИТУТ РУССКОГО ЯЗЫКА им. В. В. Виноградова Российской АКАДЕМИИ НАУК

DOI 10.15826/vopr_onom.2019.16.3.041
UDC 811.134.4’373.232 + 81:004.774.2 + 929

Xulio Sousa
Instituto da Lingua Galega
Universidade de Santiago de Compostela
Santiago de Compostela, Spain

GEONOMASTICS ON THE WEB: VISUALIZING SURNAME DISTRIBUTIONS IN A REGIONAL SPACE

The geographical distribution of surnames has been used for decades in a variety of scientific fields as a source of information for research into the history, migrations, genetics, etc. of various human populations. Nevertheless, the geographical dimension of names has hardly been explored until now in the study of anthroponyms. Situated at the intersection of human geography and linguistics, geonomastics, considered as one of the onomastic meta-disciplines, received an important impetus since the introduction of GIS technologies into humanities in the last quarter of the 20th century. This paper presents *Cartografía de Apelidos de Galicia* (CAG), a free-to-use web application developed at the University of Santiago de Compostela’s Instituto da Lingua Galega that shows the geographical distribution of Galician surnames based on the current Galician population census. The author describes the data sources covered by the app, the structure of the map, its visualizing features and search capacities. The author argues that CAG may be a useful tool for both professional and lay researchers, especially genealogists. Since most of Galician surnames derive from place names, the visualization of their geographical distribution of surnames may be insightful for identifying their origin. Complemented with historical onomastic data, this application may become in the future a valuable source of information about the migration of the population of Galicia.

K e y w o r d s: Galicia, anthroponymy, genealogy, geonomastics, surname distribution, geographic information system, digital humanities.

Software

CAG's online application was built using open source software. The J2EE platform was used to develop the webpage. The database server is in PostgreSQL with the PostGIS spatial add-on and the display screen seen by end-users is generated dynamically using the JavaScript OpenLayers library.

Acknowledgements

Cartografía dos Apelidos de Galicia is a research project of the Instituto da Lingua Galega, Universidade de Santiago de Compostela, supported by Real Academia Galega (for the 2006 version) and the Secretaría Xeral de Universidades da Xunta de Galicia (current version; TecAnDali research network, ED341D R2016/011).

The author wishes to thank César Osorio, Ana Isabel Suárez Moreno, and Dr. José Ramón Ríos Viqueira, Universidade de Santiago de Compostela, for client and server-side development insight and support. He is also grateful for the assistance received from the Real Academia Galega and the Instituto Nacional de Estadística.

1. Introduction

The study of given names and surnames is one of the main areas of onomastics. The analysis of the origin and spread of personal names, particularly surnames, has always been of great interest for other fields both within and beyond humanities [Guppy, 1890; Cheshire et al., 2009; Redmonds at al., 2015]. In the 20th century, fields as diverse as history, anthropology, human geography, linguistics and population genetics began to use surnames as a data category in research on human societies [Lloyd et al., 2004]. Geonomastics, understood as the study of patterns of territorial distribution of proper names, is considered by some scholars to be one of the most productive meta-disciplines in the onomastic sciences [Shokhenmayer, 2010; 2015; Sousa, 2017].

Progress over the last twenty years has facilitated access to large amounts of onomastic data for both professional and lay researchers, the latter often being keen on finding out about a family's origins through its surnames. Two main factors recently motivating fresh interest in onomastic documentation are the increasing accessibility of massive population databases and the rapid rise of geographic information systems using software which makes it easy to display, query, and analyse geographical data in a way that highlights patterns and trends [Béguin & Pumain, 2010; Kennett, 2012].

Early onomastic studies drew on very limited data sets obtained from heterogeneous and partial sources such as post-office directories, birth registries, and old documents. From the 1960s on, it became more common to make use of telephone directories. These name inventories are still used in some onomastic studies, but their reliability is increasingly being questioned, especially in studies of areas where the population is scattered and inadequately served by land lines. Luckily for the research community, it is becoming more usual for governmental agencies to grant scholars access to large volumes of data containing onomastic information under conditions of secrecy.

In countries such as Spain, lacking a long tradition of keeping official registers of data about the population, such censuses do not allow for the same kinds of historical comparison that are possible in more advanced countries like Great Britain and Germany [Schürer, 2004; Cheshire et al., 2009].

The development of new tools for cartographic data display and analysis has also contributed to the wider use of onomastic databases. Thanks to advances since the 1990s, GIS technology has found its way into classrooms and research laboratories as a routine work tool no longer requiring special training or heavy investment. GIS technology has also become better known due to online map applications. Onomastic studies soon began to make use of this technology, and there are now many openly accessible mapping resources that provide information about the distribution of surnames across a given geographical area [for a review see Kennett, 2012, *78–88*]. Use of these tools has resulted in products such as an English surname atlas [Barker et al., 2007], a historical dictionary of Italian surnames [Caffarelli & Marcato, 2008], and more recently, a dictionary of British and Irish family names [Hanks et al., 2016]. Maps are more than powerful visual tools that can be quickly interpreted by the audience, they also help researchers transform data into information and find distribution patterns. As Paul S. Ell pointed out, "GIS will never become an indispensible technology if it is concerned primarily with information visualization. But conceived as a tool to manage information, it will become an important humanities tool <…>. Exciting times are upon us" [Ell, 2010, *165*].

In this paper, I will present and describe *Cartografía de Apelidos de Galicia* (CAG), a free-to-use web application that shows the geographical distribution of Galician surnames based on the current Galician population census (available at http://ilg.usc.es/cag). The application's design is in line with other similar apps, incorporating some innovations which will help both specialists and amateurs interested in the origin of their surnames to make better use of the information provided. This tool came about as part of a research project on Galician surnames that was developed by Xulio Sousa and Ana Isabel Boullón Agrelo at the University of Santiago de Compostela's Instituto da Lingua Galega.

2. Data

The information that comprises CAG's corpus is drawn from Galicia's 2001 population census as furnished by the Spanish National Institute of Statistics (Instituto Nacional de Estadística, INE, see [http://www.ine.es]). The naming system used in Galicia is the same as that one employed throughout Spain, in which an individual receives two surnames: the father's first surname as their own first surname, and the mother's first surname as their second surname, e.g. ROSA CAAMAÑO VÁZQUEZ, where CAAMAÑO is the inherited father's first surname and VÁZQUEZ — the mother's first surname. In our database, each of these surnames

constitutes a separate record. The resulting database consists of 5,088,457 occurrences and 9,279 distinct surname forms attested within the Galician population which in 2001 comprised 2,695,880 individuals. Each record stores information about the municipal district in which the bearer of the name was born, the standard code of the latter and the number of people who share the same surname whether as a first or second surname. In obedience to Spanish privacy law, records of surnames occurring five times or less in a given municipality are omitted from our database. Since the surnames in the database provided by INE are given without the accent marks used in Galician and Spanish to indicate which syllable is stressed, the surnames in our database are also recorded without accents, e.g. *Fernández* and *Casás* appear as FERNANDEZ and CASAS.

Given its structure and content, the database can be queried for information about a specific surname. Entering a surname will produce a data table listing information for each territorial unit represented, in the following order: municipality, province, absolute occurrences and relative occurrences (see Table). The relative occurrence is a percentage which represents the surname's frequency in proportion to the total population of the municipality in question. This adjustment serves to compensate for the effect of heavily populated municipalities. On the results page generated by the search engine, information of interest is presented in a grid.

Results for the surname query *Sousa*

Municipality	Province	Absolute occurrences	Relative occurrences, %
Cartelle	Ourense	280	3.8499
Entrimo	Ourense	48	1.7589
Castrelo de Miño	Ourense	57	1.3933
Padrenda	Ourense	61	1.22
Lobios	Ourense	46	0.8963
Crecente	Pontevedra	44	0.7996
Mezquita, A	Ourense	20	0.6998
Pontedeva	Ourense	9	0.6787
Ramirás	Ourense	27	0.6606

3. Base map

Query results are also displayed on a map showing all the Galician municipal districts. The administrative area of the Galician Autonomous Community in north-western Spain, covering an area of 29,574 km^2, is divided into four provinces (A Coruña,

Lugo, Ourense and Pontevedra) and, from 2001 on, 315 municipalities. The division into municipalities dates from the 19th century, with the number of municipal districts fluctuating over time.

The original base map, as provided by the Instituto Geográfico Nacional, has been redrawn to improve legibility and accommodate the needs of the application. The kind of adaptations made were the usual ones for polygonal maps, simplifying linear features, eliminating details and exaggerating the size of some parts of objects [Regnauld & McMaster, 2007]. The resulting map was transferred to a shapefile format consisting of 315 polygonal figures corresponding to the Galician municipalities.

This base map allows quantitative information about the surnames in the database to be displayed. It is possible to employ different types of display which may be found useful in publications and on websites containing representations of the distribution of surnames, e.g. proportional symbols [Barker et al., 2007], dots [Munzert et al., 2014], proportional labels [Cheshire, 2011] or colours [Redmonds et al., 2015]. Given the characteristics of the data and the desire to display the information in a way which is both simple and easy to understand, we opted to represent the information by means of a choropleth map [Krygier & Wood, 2005]. This is a thematic map type by means of which quantitative data about geographical areas can be displayed effectively. In the present case, the data represented are quantitative referring to geographic areas. Differences in shading or colour are used to represent a range of values.

4. Search results page

CAG's entry page is the front page showing general information about the project and how to use the tool. A panel to the right gives general statistical information about the database together with a word cloud showing the most popular searches. The search box where a surname may be typed in is located at the top of the page. Searches may be performed on a simple string, a compound form or using wildcards. A surname search produces a page again consisting of two areas: on the left, an information panel presenting information from the database; on the right, a choropleth map which shows the distribution of surnames by municipality.

4.1. Information panel

The panel on the left shows a summary of quantitative information about the surname or search chain, including: i) occurrences: the total number of database records that coincide with the search string; ii) percent: a percentage calculated from the number of surnames found in the search result divided by the total number of surnames in the database; iii) ranking: the surname's rank in terms of its frequency out of the whole set of distinct names in the database; iv) districts: the number of municipalities in which the surname is registered.

The table beneath details the location of the surname by municipal district: name of the district, province, number of instances of registration of the surname in this district, and percentage of the total number of surnames registered in the district, where

$$p\,(\%) = \frac{number\ of\ occurrences\ of\ the\ surname\ in\ the\ district}{total\ number\ of\ all\ surname\ occurrences\ in\ the\ district} \times 100.$$

Data are sorted in decreasing order by the number in the percent column, but this order can be changed to alphabetical order by district or province or absolute numerical order (Fig. 1).

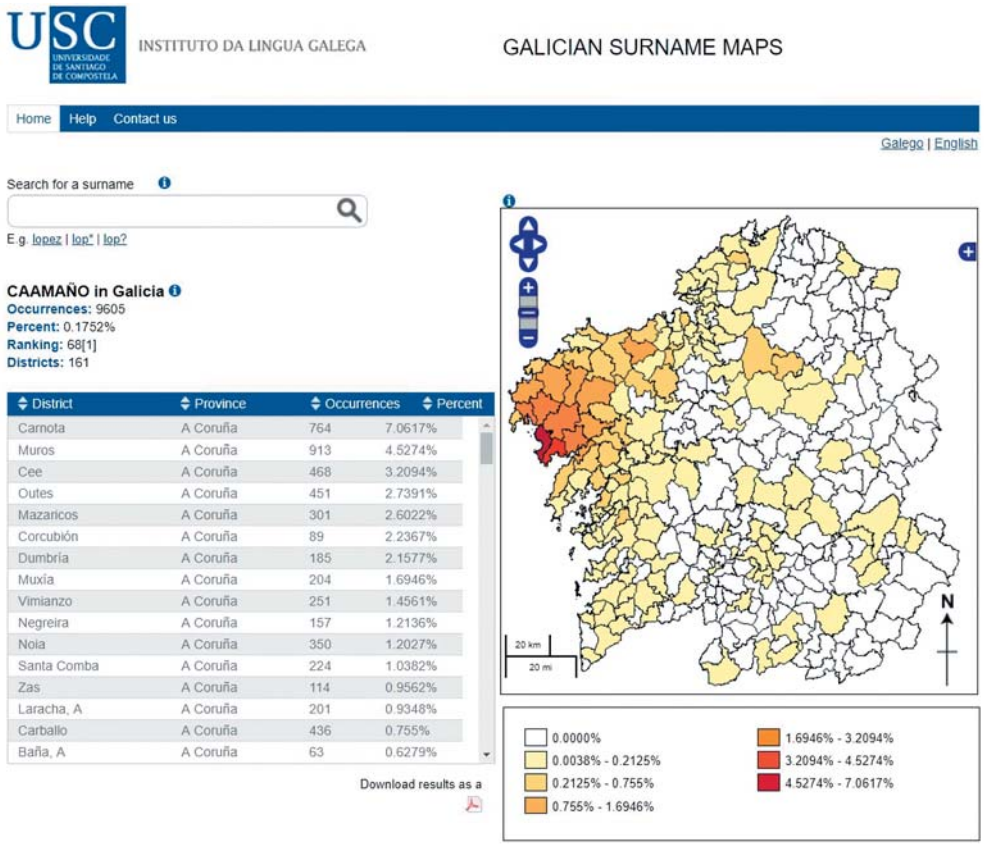

Fig. 1. Result page for the surname *CAAMAÑO*

4.2. Result map

A map displaying the search results is shown in the right-hand panel, representing the geographical distribution of the queried surname or string superimposed on the map of Galicia's 315 municipalities. To avoid a common error when producing choropleths,

instead of raw data values (total occurrences) the map shows the adjusted values, which are conveyed by means of a colour scale which groups densities into six categories at the most ranging from white (no registered records for the surname in the district in question) to maroon (districts where the highest percentages are recorded). A key is provided to indicate the meaning of each colour. The Natural Breaks method used for this classification is based on the Jenks Natural Breaks algorithm [Slocum et al., 2005, *84*].

The map comes with two tools which can be used to move the image, zoom in or out and add layers for provinces and traditional regions. Clicking on a district displays its name together with the numerical data corresponding to the district.

4.3. User response

A first version of CAG, produced in cooperation with the Centro de Supercomputación de Galicia, was launched in 2006 [Sousa, 2007]. The second version, presented here, which began testing in 2010, received substantial media coverage due to the fact that at the time it was the first map application of this type to have been developed in Spain.

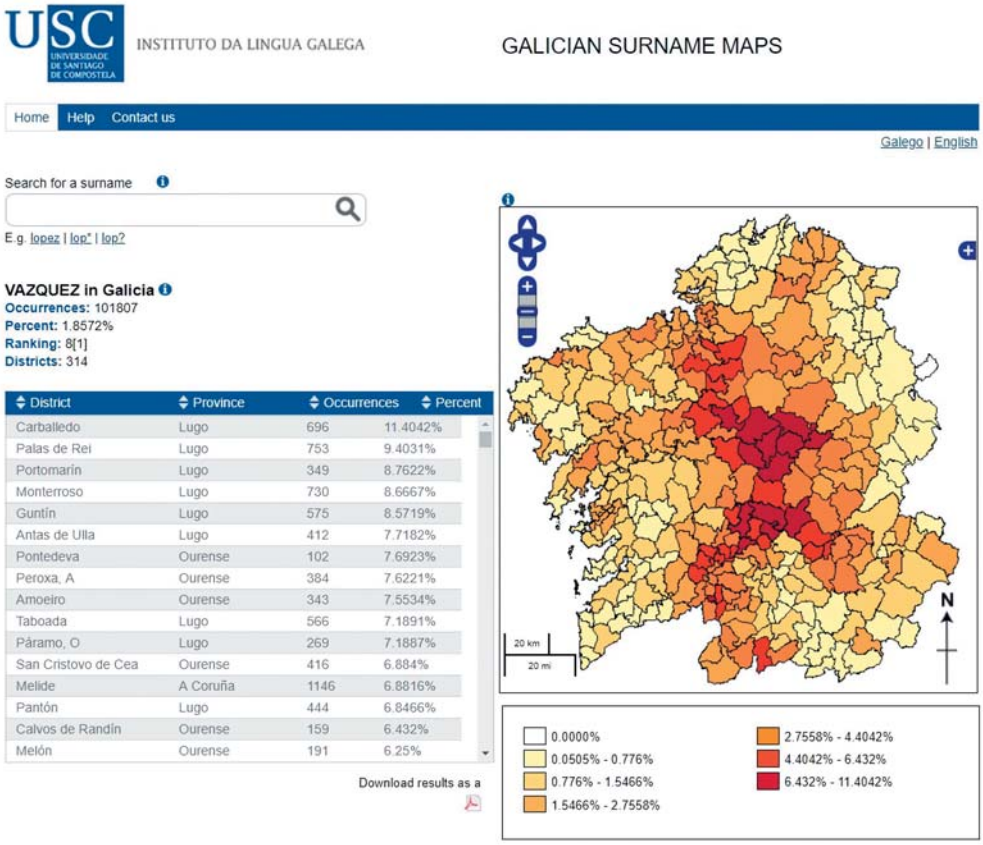

Fig. 2. Result page for the surname *VÁZQUEZ*

The Spanish INE developed and launched a similar kind of app in 2013 since when it has been continually updated and improved. Its original version and the present one, both based on data from the Continuous Register, show the distribution of Spanish given names and surnames by province. The information provided by CAG is far richer and more finely-tuned given that its point of reference is the municipal district, which is a much smaller unit than the province. CAG can be useful for those interested in researching surnames in general as well as for anybody seeking to track down the place of origin of a given surname. A great many Galician surnames originated as toponyms [Boullón, 2008]; therefore, knowledge of their geographical distribution is relevant to the question of their origin. For example, the surname CAAMAÑO (Figure 1) is more frequent in the vicinity of the place where it originated (the place called *Caamaño*). Even very frequent surnames in Galicia such as VÁZQUEZ (see Fig. 2), and indeed all patronymics ending in *-ez* (see Fig. 3), occur with significantly higher frequency in some parts of Galicia than in others.

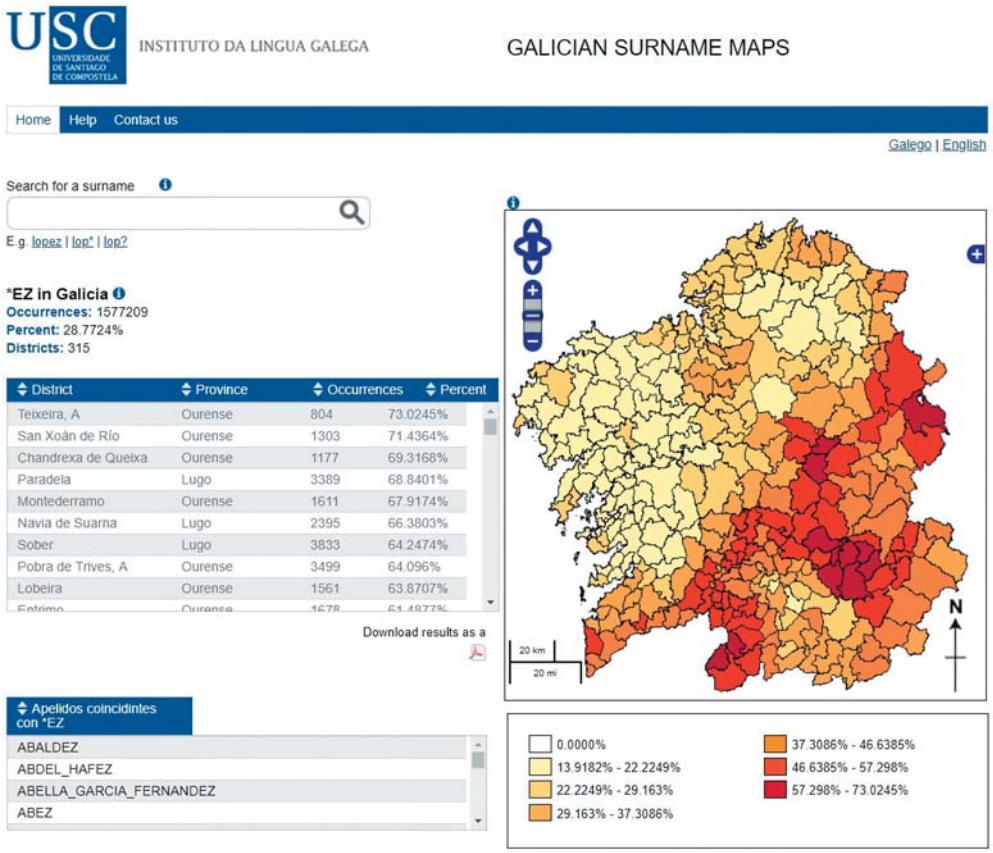

Fig. 3. Result page for the surnames ending in *-ez*

Since its launch, CAG has been the most frequently visited service on the Instituto da Lingua Galega website. The site has received over 22,800 unique visitors and there have been over 130,000 page views since 2017. The application has proved particularly popular among users in the European and American countries which hosted large numbers of Galician emigrants in centuries past.

5. Conclusions and Perspectives

Visualisation is the best way to explore and parse a large data set. Graphics and diagrams reveal the patterns covert in numerical and categorical data [Yau, 2011]. Maps are a highly intuitive and entertaining way to visualise complex data; mapping surnames brings to one's attention patterns of distribution that would be difficult to detect by analysing tables of data from a census or telephone directory without them. The CAG application has been designed as a visualisation and query tool for onomastic data generating maps containing information of interest for lay users and scholars in a variety of disciplines alike. In societies where surnames are passed on from generation to generation, the geographical distribution of surnames is not random. In a language domain such as Galicia where most surnames derive from place names, displaying onomastic information on a map is a good way to identify where names originated by observing their areas of highest density. The map that a CAG query produces gives information that may be useful in fields ranging from dialectology, history of the language and sociolinguistics to demography, human geography, history and genealogy [Hanks & Parkin, 2016].

As to our next projects, we have two goals. First, to build into the application the capacity to track the geographical spread of a given surname over time. Although no historical database is available, the population census used by CAG does include birth dates, so it would be possible to produce an animated map displaying the historical spread of a surname that might help to pinpoint its place of origin. The other goal is to develop an application similar to CAG for displaying the geographical spread of first names. Onomastic studies in several countries have already shown that first names are also distributed non-randomly [see Bloothooft & Groot, 2008; Mateos, 2014], so their study is potentially of academic interest.

Barker, S., Spoerlein, S., Vetter, T., & Viereck, W. (2007). *An Atlas of English Surnames.* Frankfurt am Main; Oxford: Peter Lang.

Béguin, M., & Pumain, D. (2010). *La représentation des données géographiques: Statistique et cartographie*. Paris: A. Colin.

Bloothooft, G., & Groot, L. (2008). Name Clustering on the Basis of Parental Preferences. *Names*, *56*, 111–163. doi: 10.1179/175622708X332851

Boullón Agrelo, A. I. (2008). Historia interna del gallego: onomástica [The Internal History of Galician: Onomastics]. In E. Gerhard, M. Gleßgen, C. Schmitt, & W. Schweickard (Eds.), *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen* (Vol. 3, pp. 3168–3177). Berlin; New York: Walter de Gruyter.

Caffarelli, E., & Marcato, C. (2008). *I cognomi d'Italia. Dizionario storico ed etimologico* [Surnames of Italia. Historial and Etymological Dictionary]. Torino: UTET.

Cheshire, J. (2011). *London Surnames*. Retrieved from http://names.mappinglondon.co.uk.

Cheshire, J., Longley, P., & Mateos, P. (2009). Combining Historic Interpretations of the Great Britain Population with Contemporary Spatial Analysis: The Case of Surnames. In *2009 5th IEEE International Conference on E-Science Workshops* (pp. 167–170). doi: 10.1109/ESCIW.2009.5407971

Cheshire, J., Mateos, P., & Longley, P. (2009). Family Names as Indicators of Britain's Regional Geography. *CASA Working Paper Series*, *149*. Retrieved from http://www.casa.ucl.ac.uk/working papers/paper149.pdf.

Ell, P. S. (2010). GIS, e-Science, and the Humanities Grid. In D. J. Bodenhamer, J. Corrigan, & T. M. Harris (Eds.), *The Spatial Humanities: GIS and the Future of Humanities Scholarship* (pp. 141–166). Bloomington: Indiana University Press.

Guppy, H. (1890). *Homes of Family Names in Britain.* London: Harrison and Sons.

Hanks, P., Coates, R., & McClure, P. (2016). *The Oxford Dictionary of Family Names in Britain and Ireland*. Oxford; New York: Oxford University Press.

Hanks, P., & Parkin, H. (2016). Family Names. In C. Hough (Ed.), *The Oxford Handbook of Names and Naming* (pp. 214–236). Oxford: Oxford University Press.

Kennett, D. (2012). *The Surnames Handbook: A Guide to Family Name Research in the 21st Century*. Stroud: History Press.

Krygier, J., & Wood, D. (2005). *Making Maps — A Visual Guide to Map Design for GIS*. New York: Guilford Press.

Lloyd, D., Webber, R., & Longley, P. (2004). *Surnames as a Quantitative Evidence Resource for the Social Sciences*. London: University College London.

Mateos, P. (2014). *Names Ethnicity and Populations. Tracing Identity in Space*. Berlin; Heidelberg: Springer. doi: 10.1007/978-3-642-45413-4

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). Mapping the Geographic Distribution of Names. In *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (pp. 380–395). Chichester: John Wiley & Sons. doi: 10.1002/9781118834732.ch15

Redmonds, G., King, T., & Hey, D. (2015). *Surnames, DNA, and Family History*. New York: Oxford University Press.

Regnauld, N., & McMaster, R. (2007). A Synoptic View of Generalisation Operators. In W. A. Mackaness, A. Ruas, & L. T. Sarjakoski (Eds.), *Generalisation of Geographic Information: Cartographic Modelling and Applications* (pp. 37–66). Amsterdam: Elsevier.

Schürer, K. (2004). Surnames and the Search for Regions. *Local Population Studies*, *72*, 50–76.

Shokhenmayer, E. (2010). Controversial Article on Geonomastics. *Meta-Carto-Semiotics. Journal for Theoretical Cartography*, *3*(1), 1–13. Retrieved from http://ojs.meta-carto-semiotics.org/index.php/mcs/article/view/29.

Shokhenmayer, E. (2015). Geography of Daily Life Names, or What is Geonomastics? In J. Tort i Donada, & M. Montagut i Montagut (Eds.), *Els noms en la vida quotidiana. Actes del XXIVè Congrés Internacional de Ciències Onomàstiques* / *Names in daily life. Proceedings of the XXIV ICOS International Congress of Onomastic Sciences* (pp. 1974–1986). Barcelona: Generalitat de Catalunya. doi: 10.2436/15.8040.01.19

Slocum, T., Kessler, F., McMaster, R., & Howard, H. (2005). *Thematic Cartography and Geovisualization*. Upper Saddle River, NJ : Pearson / Prentice Hall.

Sousa, X. (2007). Cartografía dos apelidos de Galicia: presentación do proxecto [Cartography of the Names of Galicia: Presentation of the Project]. In L. Méndez, & G. Navaza (Eds.), *Actas do I Congreso*

Internacional de Onomástica Galega "Frei Martín Sarmiento" [Proceedings of the 1st International Congress in Galician Onomastics] (pp. 327–336). Santiago de Compostela: Asociación Galega de Onomástica.

Sousa, X. (2017). Alcune riflessioni sulla geonomastica personale [Some Considerations on Personal Geonomastics]. In E. Papa, & D. Cacia (Eds.), *Di nomi e di parole. Studi in onore di Alda Rossebastiano* [On Names and Words. Studies in Honour of Alda Rosselbastiano] (pp. 387–400). Roma: SER.

Yau, N. (2011). *Visualize this*. Hoboken, NJ: Wiley.

Received 30 January 2019

* * *

Sousa, Xulio
PhD, Professor
Instituto da Lingua Galega
Universidade de Santiago de Compostela
4, Praza da Universidade
15782 Santiago de Compostela, Spain
Email: xulio.sousa@usc.es

Соуса, Шулиу
PhD, профессор
Институт галисийского языка
Университет Сантьяго де Компостела
4, Praza da Universidade
15782 Santiago de Compostela, Spain
E-mail: xulio.sousa@usc.es

Шулиу Соуса
Институт галисийского языка
Университет Саньяго де Компостела
Саньяго де Компостела, Испания

ГЕООНОМАСТИКА В СЕТИ: ВИЗУАЛИЗАЦИЯ ГЕОГРАФИЧЕСКОГО РАСПРЕДЕЛЕНИЯ ФАМИЛИЙ В ПРОСТРАНСТВЕ РЕГИОНА

Для целого ряда научных дисциплин анализ географического распределения фамилий на протяжении десятилетий служил источником информации об истории, миграции, генетике человеческих популяций. Однако до недавнего времени географическое измерение было мало востребовано в антропонимии. Находящаяся на пересечении социально-экономической географии и лингвистики, геоономастика, рассматриваемая как одна из ономастических метадисциплин, получила серьезный импульс к развитию после введения в практику гуманитарных исследований ГИС-технологий в последней четверти XX в. Автор настоящей статьи представляет геоономастический проект «Cartografía de Apelidos de Galicia» (CAG), бесплатный веб-сервис, разработанный Институтом галисийского языка Университета Сантьяго де Компостела, позволяющий визуализировать ареальную дистрибуцию фамилий Галисии, основываясь на данных последней по времени переписи населения. В статье описываются источники данных, структура карты, особенности визуального представления информации и поисковые возможности проекта. Автор

показывает, что CAG может быть полезным инструментом как для профессиональных ученых, так и для исследователей-любителей, в особенности для тех, кто занимается генеалогией. Поскольку большинство галисийских фамилий восходит к топонимам, визуализация географической дистрибуции фамилий может быть важным инструментом установления места их происхождения. Будучи дополненным историческими ономастическими данными, этот сервис в будущем может стать полезным источником информации о миграции населения внутри региона.

К л ю ч е в ы е с л о в а: Галисия, антропонимия, генеалогия, геоономастика, дистрибуция фамилий, географические информационные системы, цифровые гуманитарные науки.

Рукопись поступила в редакцию 30.01.2019