

Document downloaded from:

<http://hdl.handle.net/10251/82816>

This paper must be cited as:

Ontalba Ruipérez, JA.; Orduña Malea, E.; Alonso-Arroyo, A. (2016). Identifying institutional relationships in a geographically distributed public health system using interlinking and co-authorship methods. *Scientometrics*. 106(3):1167-1191. doi:10.1007/s11192-016-1839-z.



The final publication is available at

<https://link.springer.com/article/10.1007/s11192-016-1839-z>

Copyright Springer Verlag Germany and Akadémiai Kiadó

Additional Information

The final publication is available at Springer via <http://dx.doi.org/10.1007/s11192-016-1839-z>

Identifying institutional relationships in a geographically distributed public health system using interlinking and co-authorship methods

José-Antonio Ontalba-Ruipérez¹, Enrique Orduna-Malea^{2,*} and Adolfo Alonso-Arroyo³

¹ Instituto de Diseño y Fabricación, Polytechnic University of Valencia (UPV), 46022 Valencia, Spain.

² EC3 Research Group, Polytechnic University of Valencia (UPV), 46022 Valencia, Spain.

³ Faculty of Medicine, University of Valencia (UV), 46010 Valencia, Spain.

* enorma@upv.es

Abstract Link analysis is highly effective in detecting relationships between different institutions, relationships that are stronger the greater their geographical proximity. We therefore decided to apply an interlinking analysis to a set of geographically dispersed research entities and to compare the results with the co-authorship patterns between these institutions in order to determine how, and if, these two techniques might reveal complementary insights. We set out to study the specific sector of public health in Spain, a country with a high degree of regional autonomy. We recorded all Spanish health entities (and their corresponding URLs) that belong to, and were hyperlinked from, the national government or any of the regional governments, gathering a total of 263 URLs. After considering their suitability for web metric analysis, interlinking scores between all valid URLs were obtained. In addition, the number of co-authored articles by each pair of institutions and the total scientific output per institution were retrieved from Scopus. Both interlinking and co-authorship methods detect the existence of strength subnets of geographically distributed nodes (especially the Catalan entities) as well as their high connectivity with the main national network nodes (subnet of nodes distributed according to dependence on national government, in this case Spain). However, the resulting interlinking pattern shows a low but significant correlation ($r=0.5$) with scientific co-authorship patterns. The existence of institutions that are strongly interlinked but with limited scientific collaboration (and vice versa) reveals that links within this network are not accurately reflecting existing scientific collaborations, due to inconsistent web content development.

Keywords Link analysis; Network analysis; Co-authorship; Public health; Institutional relationships; Spain.

1. Introduction

The structure and intensity of online relationships may be examined using different approaches (Barnett and Park 2014). For example, we can consider the physical network (measuring the bilateral internet bandwidth), the content network (measuring the web traffic from specific websites) or the navigational network (measuring the hyperlink connections among web domains). All these techniques make it possible to arrive at an approximation of phenomena that occur in the offline world from phenomena that occur in the online world. Among these different methods, we can highlight the development of link analysis (Thelwall 2004).

Since its inception over a decade ago, link analysis has been widely applied to the study of relationships between organisations on the web (Park and Thelwall 2003; Thelwall 2009), during which time a constantly evolving theoretical framework has been constructed (Thelwall 2006; Orduna-Malea and Aguillo 2014). The rationale behind this method is that relationships between organisations can be approximately quantified through the relationships between the URLs that represent them; so the organisation of the web acts as a mirror of social structure (González-Bailon 2009). Therefore, the construction of networks by means of URL mentions (when the URL that corresponds to website B is mentioned on website A, whether a hyperlink is activated or not) may reflect either expected or unexpected relationships among institutions.

Several authors have used link metrics to map the volume of links exchanged between two or more websites (Lang et al. 2014), estimate the level of collaboration between institutions (Stuart and Thelwall 2006), describe the structure and topology of these navigational relations (Ortega and Aguillo 2009) or even correlate the volume of links received with other indicators, such as citations, in the case of research institutions (Almind and Ingwersen 1998; Thelwall 2001). These analyses may provide information reliable enough to be used for policymaking (Thelwall, 2002).

Likewise, hyperlink analyses have been applied to a wide set of environments, which include, among others, business (Vaughan and Wu 2004; Vaughan and You 2006; Orduna-Malea et al. 2015), industry (Stuart and Thelwall 2006; Romero-Frias and Vaughan 2010a; Orduna-Malea and Aytac 2015), and politics (Park 2011; Romero-Frias and Vaughan 2010b). However, these are open areas where link analysis becomes more complicated and encounters both conceptual and methodological problems.

One reason for link analysis being somewhat problematic is that linking has a very low cost barrier (Weber and Monge 2011). Therefore, hyperlink analysis may contain irrelevant information or inconsequential relationships (Barnett and Park 2014). Another is that, if these analyses rely on the use of commercial search engines, we can find a number of technical limitations, especially in the accuracy and precision of web metrics based on hit count estimates (Vaughan and Thelwall 2004; Wilkinson and Thelwall 2013).

The limitations of link analysis have led to its use expanding more in certain academic and controlled environments – such as universities (Aguillo et al. 2006) and academic journals (Vaughan and Thelwall 2003) – than in more open and uncontrolled environments – such as private companies, organisations or public institutions. Moreover, if link metrics are used for relational purposes (to determine whether the relationship between “web domain a” and “web domain b” is stronger than that between “web domain a” and “web domain c”, for example), these limitations are minimised since all URLs are subject to the same error, and thus the error is statistically dispersed.

Nevertheless, and beyond all the well-known advantages and disadvantages, link analysis possesses a special feature: its high sensitivity to cultural, linguistic and geographical differences, unattainable with bibliometric indicators. For example, Thelwall (2002) found that interlinking among universities in a homogeneous country (UK) decreases with distance while Vaughan (2006), who examined how linguistic and cultural factors affect university relationships in a heterogeneous country (Canada) through co-links, concluded that views on French Canada are based more on linguistic or cultural difference than geographical location.

In the case of Europe, despite the finding of Thelwall et al. (2003) that English is the standard web language for linking throughout the EU, Heimeriks et al. (2003) detected cultural and linguistic patterns among the relationships of 220 European universities. Moreover, in the case of national relationships, different behaviour was detected in comparison with international relationships. Ortega (2007) found that the majority of Western European academic websites are international and multilingual, with English and national languages operating in tandem throughout, a fact that explains why English is influencing and attracting more external links to European universities than other

languages (Orduna-Malea et al. 2014). Nonetheless, the influence of English may be different in other countries outside Europe (Thelwall and Tang 2003).

Notwithstanding, this sensitivity property makes link analysis a useful tool as a complement to bibliometric studies oriented to the analysis of relationships among institutions (mainly through co-authorship patterns). Thus it makes sense to assume that if two institutions do collaborate closely (by means of a number of articles published jointly) this fact may mean that some relationship between these institutions actually exists. And this relationship should, to a certain extent, be reflected on the web if the corresponding institutional websites are designed adequately and if they are suited to webometric analysis.

Nonetheless, the underlying human motivations for link creation – even within such specific and controlled spaces as research institutions – are complex, and professional, research-oriented and informative issues seem to be the main reasons for link creation among these websites (Wilkinson et al. 2003; Bar-Ilan 2005; Seeber et al. 2012). This means that hyperlinks and co-authorship data may reflect both different and complementary aspects of institutional relationships. However, the correlation between these two metrics should be statistically significant if institution websites actually reflect the scientific activities of the corresponding institutions, either directly (openly depositing academic material) and/or indirectly (informing about academic undertakings).

The complementarity of these two institutional relationship views is of special interest when studying systems composed of institutions that are related but with different purposes or activities (i.e., governmental, research, informative or management, or all of these activities combined but with different intensity), where geographical differences among institutions can also determine the degree and nature of their relationships. This is the case for studies centred on Triple-Helix relationships (THR). Specifically, THR analyses of an entire heterogeneous country will be strongly determined by its legal and political structure (more centralised or more distributed) as well as by geographical and language patterns (if different official local languages co-exist, for example).

Triple Helix studies could thus benefit from link analysis techniques to achieve a better understanding of the relationship between different institutions (Khan and Park 2011), offering alternatives to strictly coded and commercially available indicators like co-authorship publications. The application of web metrics to THR has, however, been scarce to date (Leydesdorff and Curran 2000) and applied to excessively aggregated entities with few exceptions (Stuart and Thelwall 2006; Garcia-Santiago and Moya-Anegón 2009). Heimeriks and Van den Besselaar (2006) advise of the need to disaggregate data obtained through link analyses from complex institutions – such as universities – since link patterns depend on scientific fields, so the analysis of lower-level, disaggregated institutions might offer more detailed information.

Therefore we have found a gap in the literature on the analysis of Triple-Helix relationships: taking geographical differences into account and combining hyperlinks and publications in order to achieve a better understanding of the relationships among disaggregated institutions. Moreover, in contrast to the University and Industry Helices, government linking (the analysis of links between government websites or websites

depending on public government, including research bodies) has hardly been studied at all, despite its considerable interest, with the exception of works by Holmberg (2010) and Holmberg and Thelwall (2009), although these studies apply to homogenous government entities in very local environments with results that are difficult to extrapolate. But a nationwide study on a set of heterogeneous entities depending on public government has not yet been carried out.

2. Objectives

The main objective of this study is to determine how, and if, the combination of link analysis and co-authorship techniques can reveal complementary insights when identifying the institutional relationships within a geographically distributed system composed of a heterogeneous set of public government entities (including research bodies).

To accomplish this objective, the following secondary goals were set out:

- a) Determine the suitability of the analysed entities for link analysis (mainly by URL syntax).
- b) Study the web impact of the corresponding institutions (mainly by page count and web visibility metrics).
- c) Analyse the interlinking patterns between institution web domains.
- d) Ascertain the correlation between interlinking and co-authorship data.

The first two secondary objectives were necessary prior to the interlinking analysis to characterise the set of web domains studied, since the URL syntax used by institutions and the amount of content (i.e. the relationship between content and the total number of links received) may help to explain the specific number of hyperlinks created between institutions.

3. Method

Selecting the sample

The sample of institutions was obtained by studying a specific sector (public health) in a specific country (Spain).

On the one hand, the public health sector constitutes a very special and specific case study within the Government Helix since it is a complex system that brings together a wide range of entities (research institutions, health centres, government bodies, etc.): it provides different relationship patterns for link analysis and co-authorship techniques. On the other hand, Spain is a particularly apposite research subject because it is a nation with a high degree of regional autonomy, large cultural differences and diverse co-official local languages, which means greater complexity in the design and analysis of distributed system data which in turn could determine institutional relationships.

Although the Spanish public health system has been partially analysed from a bibliometric point of view, with special mention of the Spanish bibliometric map in Biomedicine and Health Sciences (Méndez-Vasquez et al. 2008), currently updated to

the period 1997-2011 (Méndez-Vásquez et al. 2012), data on co-authorship between all Spanish health system institutions is lacking. Moreover, García-Lacalle et al. (2011) additionally detected an absence of studies on healthcare web presence, amply justifying the choice of this sample for this study.

The political distribution of health authorities in Spain is regulated by: the Spanish Constitution of 1978; General Health Act 14/1986 of 25 April; and Act 16/2003 of 28 May on cohesion and quality in the National Health System (Ministry of Health, Social Services and Equality 2012). According to this distribution, the national government is responsible for the overall coordination of national healthcare, foreign healthcare, and drug legislation, while the regions are responsible for health planning, public health and the management of health services.

Obtaining the sample

In the first phase of the study we recorded all Spanish health entities that belong to the national government or to any of the regional governments. The regions comprise the 17 autonomous communities and two autonomous cities into which Spain is divided (see Appendix I), and correspond to Eurostat NUTS 2 level (2011). We included all public and mixed, but not private, entities.

Hospitals have been excluded from this study as it would have been extremely complicated to include them. These establishments (that may depend directly on public universities, national or regional government, private entities or any combination thereof) currently have a web presence that reflects hierarchies and institutional relationships poorly in their corresponding official URLs and where the existence of consortia further complicates web analysis. Because of this, hospitals are being studied independently. Since this study has a more methodology-based approach, the exclusion of these establishments does not affect the proposed objectives: the combination of interlinking and co-authorship analysis methods to identify institutional relations in a distributed national public sector, rather than a mere description of the sector itself.

Next, the corresponding URL of each institution was obtained. We established that the unit of study would be the more general URL of the entity website (Orduna-Malea and Ontalba-Ruipérez 2013), as this contains all the web pages of the corresponding institution. All entities whose website contained a single web file or a limited number of web files under different URL hierarchies (different subdomains or subdirectories) were excluded since this web practice jeopardises link analysis. Unfriendly URLs (those including diacritics or non-optimised dynamic URLs) were excluded as well since search engines cannot retrieve web metrics accurately from them.

The identification and listing of Spanish public health institutions (and their corresponding URLs) was performed manually by browsing and searching all official websites related to health at national and regional level. The process was as follows: we began by browsing the different official national (Ministries) and regional websites (Councils), searching for all hypertext references to any element of the Spanish health system on all levels. The URLs that were identified and validated for these websites were subsequently revisited to repeat the process of identifying all hypertext references to any element of the Spanish health system.

If a health service, research centre or any other institution was not linked to from these seed websites, it was discarded. This process was conducted between June and September 2014.

Obtaining web metrics

Next we obtained the following web impact indicators: page count (number of files hosted within a domain name and indexed in search engines) and web visibility (number of links or sites linking to a website) only for those URLs whose syntax allows these data to be captured. This enabled us to determine not only the suitability of the Spanish online public health system for link analysis but also its impact. Since the different parts of a URL are not always used consistently, the nomenclature used in describing URL syntaxes is described in Figure 1

Figure 1. URL syntax nomenclature

The indicators used, their scope and the source used to obtain them are summarised in Table 1 (web impact metrics). In the case of web page count, the source used was Google (<google.com>) because it is the most comprehensive search engine in the world today. Since the Spanish version (<google.es>) showed similar results for a small sample of data, the international version was used.

Use of the page count indicator can generate noisy data because it includes all kinds of material (images, icons, multimedia, etc.). For this reason, file type filtering may be necessary in some cases to obtain cleaner data. However, filtering also gives rise to various problems:

- First, different web content generation practices prevent us from knowing a priori the best file type to filter. Some websites may have few HTML files but a lot of rich PDF files or collections of images of interest; conversely, other websites may have many banal HTML files.
- Secondly, there is an increasingly widespread practice of hiding extensions in URLs, which means that filtered file quantification cannot be performed with precision for certain file types (HTML in particular).
- Third, web files also present additional problems due to the already widespread use of Content Management Systems (CMS) that generate dynamic URLs without a file type extension.
- Finally, automatic searches through query commands depend on the performance of advanced search (in this case Google). For example, it is no longer possible to search for HTML files directly through the “filetype” command.

Since total page count correlates highly with the specific page count of rich files (Orduna-Malea et al. 2014), we decided to use total page count for data correlation despite the noise it generates. In any case, PDF page count was also used to check correlation with overall data size.

For web visibility data, it was not possible to obtain domain-level link indicators with Google (Thelwall and Sud 2011) so the source used was Open Site Explorer (OSE),¹ an online application created in 2010 to extract and analyse link data from the Mozscape

Index, a data repository that provides access to over 145 billion URLs (12 November 2015). It also provides free and paid services, search engine or API requests, and generates its own metrics (MozRank, MozTrust and Domain Authority). OSE can be used as a reliable tool for collecting webometric data, as previous research has demonstrated (Orduna-Malea et al. 2015).

Table 1. Web indicators: impact (page count and visibility) and network

Web impact metric	SCOPE	SOURCE	QUERY
Page count	Files indexed within a web domain	Google	site:url.tld
PDF page count	Files indexed within a web domain in PDF file type		site:url.tld filetype:pdf
Total external links	Links to a website that come from pages not on the same root domain	Open Site Explorer	Direct
Total linking root domain	Unique root domains containing at least one link to a specific URL	Open Site Explorer	Direct
Domain MozRank	Popularity metric (on a logarithmic scale from 1-10) similar to PageRank	Open Site Explorer	Direct
URL mention	Times a URL is mentioned from a specific root domain	Google	“abc.tld” site:xyz.tld
Node-level metrics	SCOPE		
Degree	Number of edges that are adjacent to one node		
Betweenness	The average distance from a given node to all other nodes in the network		
Clustering coefficient	The degree to which nodes of the neighbourhood of a node “a” are connected to each other		
Network-level metrics	SCOPE		
Average degree	The average degree over all of the nodes in the network		
Average clustering coefficient	The average clustering coefficient over all of the nodes in the network		
Average path length	The average graph-distance between all pairs of nodes		
Diameter	The maximal distance between all pairs of nodes		
Density	How close the network is to complete (density equal to 1)		

Finally, all data were entered into a spreadsheet, where they were subjected to statistical analysis with the XLSTAT application. We obtained the correlation coefficient between the impact indicators to ascertain the possible influence of web content on total links received (an issue that may determine interlinking behaviour). The Spearman coefficient was applied, in light of the generally unequal distribution of web data (Barabasi and Albert 1999). A Principal Component Analysis (PCA) was also conducted in order to visualise the different dimensions of the web indicators used. This statistical analysis was carried out between September and October 2014.

Obtaining interlinking metrics

In the case of selective links (links from a particular domain to another) neither Google nor Open Site Explorer may be used, at present, to obtain this indicator (Sud and Thelwall 2014). Therefore a proxy created from the URL mentions indicator was employed, a method already used in scholarly literature (Ortega et al. 2014), using Google as a data source.

Queries were made by introducing the following syntax in the search field: <“abc.tld” site:xyz.tld>. This query returns the number of times the domain “abc.tld” has been

mentioned in the domain <xyz.tld> and is therefore an indicator of the intensity of the relationship between the two web domains.

We thus created all the possible combinations between all the URLs recorded in the previous phase, whether they were domains or subdomains (syntax types where visibility indicators function correctly). For each combination, the hit count estimate provided by Google was obtained manually (this task was performed equally by all the authors in order to check and classify possible errors with URL mentions; this task took four weeks). Subsequently, a .NET file was generated manually from these data and imported into the Gephi application, which was used to produce a graph of all the relationships.

The same application was used to obtain various indicators (Table 1) at both node (Degree, Betweenness, Clustering Coefficient) and network level (Average Degree, Average Clustering Coefficient, Average Path Length, Diameter, Density, Eigenvector); these indicators describe the characteristics of the network. Impact indicators and URL mentions were recorded between October and November 2014.

Obtaining co-authorship metrics

The total scientific output for each institution (only those entities with valid URLs for link analysis) was retrieved from Scopus. Then, the identification (ID) provided by Scopus for each institution was used to obtain the number of publications co-authored by each pair of institutions through the following query search: <AF-ID(institution1) and AF-ID(institution2)>. These bibliometric data were retrieved on July 2015, and the total cumulative data at this time for each institution were retrieved. Given the slower pace of the publication process compared to the creation of hyperlinks, we consider the gap of about ten months between the retrieval of hyperlink and co-authorship data to be useful for the analysis of academic relationships that already existed when the link data were retrieved.

Finally, the correlation between interlinking and co-authorship data was obtained. Since this information was compiled in two different matrices, the Mantel test (Spearman; two-tailed test) was performed. In view of the fact that interlinking is a non-reciprocal process (the number of links from institution A to institution B is not equal to the number of links from institution B to institution A), all links between each pair of institution URLs were added up prior to the correlation process. In the case of institutions with alternative web domains (more than one valid URL), mentions to/from each of the alternative web domains were considered. This statistical analysis was carried out during July 2015.

4. Results

4.1. Suitability of Spanish online public health system for link analysis

A total of 263 URLs were located, of which 25.1% (66 URLs) are national, while the remaining 74.9% (197) belong to the 19 regions, of which Catalonia, with 17.5% (46), Andalusia, with 8.7% (23) and the Canary Islands, with 7.6% (20), stand out in particular. Appendix II (see complementary material) shows all analysed URLs,

grouped by area of jurisdiction (national or regional), and URL type (domain, subdomain, subdirectory, combined).²

With regard to URL type, domain is the most used syntax, accounting for 46.7% (123 URLs) of all the websites analysed. Nationally it accounts for 62.1% of URLs. Regionally, the average percentage of domains is 41.6%, Catalonia leading the category with domains making up 71.7% (33) of the total for that region, while for Andalusia, despite it being the region with the second-highest total number of domains (10), these domains make up only 43.5% of the total. Finally, it is worth noting that Madrid does not have any domains. The subdirectory is the second most used syntax, with 82 URLs (31.2%); the Canary Islands and Andalusia are the standout regions, with 15 and 13 subdirectories respectively. Finally, the subdomain, with 41 URLs (15.6%), is the third most used syntax, the region of Valencia having the highest number, with 7 web addresses. Additionally, 17 URLs (8 national and 9 regional) composed of a combination of domain/subdomain and subdirectories were identified.

Moreover, subdomains, subdirectories and combined syntax may be used at different depth levels, thus reflecting a particular knowledge organisation. For each of the 263 URLs located, both their depth and their degree of friendliness were analysed. Full details may be consulted in the complementary material (Appendix IV).

The general distribution of URL syntaxes according to web depth level is illustrated in Table 2. At regional level, we encountered greater intensity in the design of subdirectories (38 URLs presented more than two hierarchical levels), which greatly hinders their use for not only web metric purposes but also search engine optimisation.

Table 2. Distribution of URL syntaxes and web depth levels

Scope	Next level	Subdomain 2 nd level	Subdomain 1 st level	Domain	Subdirectory 1 st level	Subdirectory 2 nd level	Next level	Total
National	0	6	8	41	1	0	0	56
Regional	0	6	20	82	20	22	38	188
Total	0	12	28	123	21	22	38	244

Combined URL syntaxes have been excluded

With regard to combined syntax, this URL type was concentrated at national level in the Ministry of Health, Social Services and Equality (specifically in subdirectories within <msssi.gov.es>), and at regional level in the Basque Country (5 URLs) and Valencia (3).

Furthermore, unfriendly URL types (i.e. dynamic URLs that have not been optimised for search engines) were detected in both subdirectories and combined syntax. This effect was observed in 38 URLs (14.4% of all URLs analysed), all at regional level; it is particularly significant in the case of the Madrid Regional Ministry and the Regional Ministry of Health and Social Services of La Rioja when we consider that they are high-level administrative structures.

Finally, of the initial 263 URLs, 38 unfriendly URLs were discarded, as were two other URLs (<ib-salut.caib.es> and <osakidetza.euskadi.net/r85->) that are not websites per se, rather simple web pages. This produced a total of 223 valid URLs for web impact analysis, divided into 131 domains, 44 subdomains and 48 subdirectories.

4.2. Web impact of Spanish online public health system

Web impact metrics

Table 3 shows the sites whose URL contained more than 100,000 files (a total of 13 centres), with the National Centre for Biotechnology ranked first (page count of 850,000), followed far behind by the Spanish Agency for Medicines and Health Products (386,000) and the Carlos III Health Institute (376,000); all of these are national centres. Conversely, low values (a page count of fewer than 100) were obtained in 54 URLs (24.21%), of which 26 were subdirectories; in 6 (2.7%) no result was obtained. Full data is available in Appendix V.

Table 3. Top health centres according to page count with web visibility metrics

ENTITY	PLACE	PAGE COUNT	WEB VISIBILITY		
			MzRank	TEL	TLRD
National Centre for Biotechnology	National	850,000	2.58	2	3
Spanish Agency for Medicines and Health Products	National	386,000	1.64	14	7
Carlos III Health Institute	National	376,000	5.76	102,762	3,473
Regional Health Ministry	Region - Valencia	373,000	3.37	7	6
Health Service	Region - Murcia	371,000	3.42	2,943	20
Germans Trias i Pujol Health Sciences Research Institute	Region - Catalonia	327,000	4.17	34,548	74
Hospital del Mar Medical Research Institute	Region - Catalonia	239,000	5.31	110,268	841
Galician Health Service	Region - Galicia	230,000	5.19	40,140	1,132
Ministry of Health, Social Services and Equality	National	185,000	3.33	374	19
Valencian School for Health Studies	Region - Valencia	168,000	3.7	9	7
National Cancer Research Centre	National	160,000	5.48	18,324	1,046
Biomedical Research Institute of Lleida	Region - Catalonia	155,000	4.22	503	64
Biological Research Centre	National	140,000	5.06	2,132	282

TEL: Total external links; TLRD: Total Linking Root Domains

To analyse web visibility, subdirectories were discarded (since link indicators are not precise enough for this URL type), which left us with 175 URLs (domains and subdomains), from which the total number of external links, root domains and MozRank were extracted.

We observed an uneven distribution of external link data. The figure of 1,000 links was attained by more than 70 entities (40%), a figure that is by no means negligible (the Hospital del Mar Medical Research Institute, which ranks first for this indicator, received 110,268 links, followed by the Carlos III Health Institute with 102,762). On the other hand, we identified 29 entities which received fewer than 50 external links (details in Appendix V), with the National Centre for Biotechnology faring particularly poorly (only 2 external links); as mentioned above, this is the entity with the largest web

page count. Finally, special mention must be made of the high web visibility of the Catalan centres (in fact, of the 50 centres with the most external links received, 18 pertained to Catalonia, while 15 were national).

However, this high visibility diminishes when the links are aggregated per linking root domain. For example, the National Plan on Drugs (<pnsd.msssi.gob.es>), despite receiving 15,793 external links, received the links from only 102 different websites. We noted 18 centres for which the number of external links per root domain exceeded 100 (see Appendix VI). This indicates that only a few sites provide the bulk of the external links. For this reason the usefulness of external links is limited in this regard, and it is preferable to use alternative document models such as linking root domains (LRD) to more adequately discern the web impact of units.

Linking root domains also show an uneven distribution of data; while 60 entities (34.3%) had over 100 root domains, 40 (22.9%) did not attain the figure of 20 root domains, which is considered a low figure given the importance of the centres analysed. The Andalusian School of Public Health (3,878), Carlos III Health Institute (3,473), Cantabrian Health Service (2,131), Galician Health Service (1,132) and the National Cancer Research Centre (1,046) are the institutions that attained the highest number of root domains.

An even more elaborate step is to quantify the quality of the sites from which the links are received. In this aspect, the data provided by MozRank are generally very scarce. Only 14 of the 175 URLs analysed scored higher than a 5 (maximum score 10): the Carlos III Health Institute (5.76) obtained the highest score, followed by the National Cancer Research Centre (5.48).

Correlation between web impact metrics

To verify the possible association or dissociation between page count and web visibility metrics, the correlation between all web impact indicators was calculated (Table 4).

Table 4. Spearman correlation between web impact indicators

November 2014	Total Page count	MozRank	TEL	TLRD
Total page count	1.00			
MozRank	**0.42	1.00		
TEL	**0.46	**0.70	1.00	
TLRD	**0.47	**0.87	**0.79	1.00

November 2015	Total Page count	PDF Page count	MozRank	TEL	TLRD
Total page count	1.00				
PDF page count	**0.75	1.00			
MozRank	**0.40	**0.39	1.00		
TEL	**0.47	**0.48	**0.76	1.00	
TLRD	**0.49	**0.50	**0.87	**0.88	1.00

** Significant values (except diagonal) at the level of significance alpha=0.010 (two-tailed test)

TEL: Total external links; TLRD: Total linking root domains.

Although all the correlations were statistically significant ($\alpha=0.01$), page count values were slightly lower than expected for MozRank ($r=0.42$), total external links ($r=0.46$) and root domains ($r=0.47$). On the other hand, high correlation levels were obtained between different visibility data, especially between root domains and MozRank ($r=0.87$).

All correlation data were recalculated a year later (November 2015) in order to test their reliability and to run a control on the PDF page count (HTML page count was omitted for the technical reasons outlined above). Results show that the correlations obtained a year later are very similar, and maintain a clear separation between size and visibility indicators. The correlation between total page count (2014) and total page count (2015) is also very high ($r=0.866$), reinforcing the validity of the data obtained.

In order to visualise the relationships between indicators for the sample of units under analysis, a principal component analysis (PCA) was conducted in which the unit level (national or regional) was included as a qualitative variable to detect any patterns in the performance of the indicators based on this variable (Figure 2).

Figure 2. Principal Component Analysis (PCA) of units according to region

Figure 2 shows the separation between the page count variable and the three visibility measurements, where the first two factors of analysis explain 89.6% of variance, a high value which reinforces the separation of these two indicator dimensions.

Moreover, we observe the existence of outliers, especially in national units, both in page count (National Centre for Biotechnology, <cnb.csic.es>; Spanish Agency for Medicines and Health Products <aemps.gob.es>) and web visibility metrics (Carlos III Health Institute, <isciii.es>). Similarly, values for some regional centres also produced outliers, both in page count, e.g. the Valencia Regional Health Ministry (<san.gva.es>) or the Murcia Health Service (<murciasalud.es>), and visibility (Andalusian School of Public Health, <easp.es> or the Hospital del Mar Medical Research Institute, <imim.es>), as discussed above.

4.3. Interlinking patterns between Spanish online public health institutions

Finally, URL mentions from each URL to the others (a total of 30,450 combinations) were calculated from the final sample of 175 URLs corresponding only to domains and subdomains to create a directed graph of the composition of the Spanish health system on the Web.

Characteristics of the network

The results show a network with very low density values ($D=0.06$), average degree (10.05) and average clustering coefficient (0.41), while values were relatively high for diameter (6) and average path length (2.35).

Low density values were determined by the low relationship intensity (i.e. the low number of mentions from one URL to another). In fact, 94.2% of the combinations (28,692) gave a null value while more than 100 mentions were only recorded for 72

combinations (the combinations with the highest intensity in the network are available in Appendix VII). It shows how relationships of greater intensity are linked to the website of the Ministry of Health, Social Services and Equality (<msssi.gob.es>), and in particular to the relationships between entities hosted on subdomains within the Ministry.

In any case, these nodes, despite belonging to the same entity, have been kept separate in the network in order to analyse the influence of each URL separately. Table 5 lists the 20 URLs with the greatest betweenness centrality, with the Carlos III Health Institute (8,617.08) in first position, followed at some distance by the Ministry of Health, Social Services and Equality (3,470.93). Further details about network metrics are provided in Appendix VIII.

Table 5. Node-level network indicators

URL	In Degree	Out Degree	Degree	Betweenness	Clustering Coefficient	Eigenvector
isciii.es	89	109	198	8,617.08	0.08	1.00
msssi.gob.es	46	83	129	3,470.93	0.11	0.58
sergas.es	50	34	84	1,693.94	0.17	0.67
san.gva.es	32	39	71	996.57	0.20	0.45
ont.es	24	14	38	972.53	0.23	0.35
aemps.gob.es	32	39	71	924.68	0.15	0.43
ciber-bbn.es	23	38	61	666.60	0.20	0.36
idibaps.org	19	32	51	656.37	0.18	0.27
prbb.org	25	46	71	625.42	0.19	0.39
cnb.csic.es	24	43	67	623.84	0.20	0.40
saludextremadura.com	16	25	41	536.13	0.25	0.26
ciberer.es	28	37	65	529.60	0.27	0.45
bioef.org	15	19	34	514.54	0.28	0.25
cnio.es	40	26	66	476.50	0.22	0.59
riojasalud.es	19	9	28	467.54	0.37	0.29
irbbarcelona.org	21	30	51	459.48	0.26	0.30
murciasalud.es	20	37	57	392.12	0.24	0.33
easp.es	26	18	44	388.75	0.32	0.38
iacs.aragon.es	28	30	58	366.51	0.27	0.42
fimabis.org	7	6	13	366.26	0.18	0.03

Figure 3 shows the online network of the sampled institutions. High centrality may be observed in the nodes corresponding to national units (in red) where, as well as the Ministry of Health, Social Services and Equality (<msssi.gob.es>) and the Carlos III Health Institute (<isciii.es>), other major centres also feature, such as the Spanish Agency for Medicines and Health Products (<aemps.gob.es>), the National Centre for Biotechnology (<cnb.csic.es>) and the National Cancer Research Centre (<cnio.es>). Next to this central zone, dominated by national units, are the regional health centres of Galicia (<sergas.es>), Basque Country (<osakidetza.euskadi.net>), Murcia (<murciasalud.es>) and Valencia (<san.gva.es>), all of which are close to the Ministry node.

Figure 3. Spanish online network (n=175; Fruchterman Reingold algorithm)

Around the central zone, containing the national units, the regional units with the highest number of nodes in the network are located: Catalonia (yellow), Andalusia (green) and Extremadura (brown), with a clear cohesion between them. Valencia (blue),

on the other hand, is much more widely dispersed. The other nodes, regardless of the region to which they belong (and with very few nodes in the network) appear in grey.

Hierarchical relationships

Data for relationships between the various institutions under the Ministry of Health, Social Services and Equality are highly inflated due to both the institutional relationship and a shared parent domain. For example, from <msssi.gob.es>, 73,100 hits to <pnsd.msssi.gob.es> were recorded, 56,500 to <aesan.msssi.gob.es> and 9,360 to <ingesa.msssi.gob.es>.

Subnets

Given the complexity of the overall network, and the evident geographical influence, we disaggregated data for the national subnet (Figure 4, top), and the regional subnets of Catalonia (Figure 4, centre left) and Andalusia (Figure 4, centre right), which comprise the highest number of nodes (38 and 18 respectively).

Figure 4. National and Regional (Catalonia and Andalusia) networks and metrics.

These subnets also obtained very low density values. The national network (formed by 55 nodes), has a density of 0.126 (Figure 4, bottom), while that of Catalonia, which includes the relationship with the highest intensity in the whole network (that between Hospital del Mar Medical Research Institute <imim.es> and the Barcelona Biomedical Research Park <prbb.org>, with 15,400 mentions) is slightly higher (0.148).

Furthermore, the two most complex subnets (national and Catalan) are interconnected, due mainly to the relationships between the nodes with the highest eigenvector value in both networks (<isciii.es>, <cnio.es> and <msssi.gob.es> in the national network, and <imim.es>, <iconcologia.net>, <prbb.org> and <ibecbarcelona.eu> in the Catalan network), as can be seen in Figure 5.

Figure 5. Interconnection of the national and Catalan subnets

Asymmetry

It should be noted that the relationships established in the different networks and subnets are directed and, therefore, not symmetrical. Moreover, the high average degree for some nodes is sometimes due to the high number of mentions towards the other nodes of the network (e.g. the Ministry of Health, Social Services and Equality, which acts as a Hub) or to the number of mentions received from other nodes (e.g. the Hospital del Mar Institute of Medical Research, which acts as an Authority). Table 6 shows the main Hub and Authority nodes in the network.

Table 6. Main Authority and Hub nodes in the network

HUBS			AUTHORITIES		
URL	Av Degree	Asymmetry	URL	Av Degree	Asymmetry
msssi.gob.es	129	-37	imim.es	42	42
cibersam.es	51	-25	cnic.es	44	26

rticc.org	36	-24	iconcologia.net	39	21
prbb.org	71	-21	sergas.es	84	16
cerca.cat	31	-21	cabimer.es	19	15
isciii.es	198	-20	imppc.org	15	15
cnb.csic.es	67	-19	cnio.es	66	14
opimec.org	28	-18	pasteur.crg.es	14	14
murciasalud.es	57	-17	cib.csic.es	27	13
redheracles.net	27	-17	aatrm.net	13	13
fundacioncadiz.es	17	-17	saludcantabria.org	14	12
ciber-bbn.es	61	-15	predimed.es	14	12

Asymmetry: InDegree – Out Degree; Av Degree: InDegree + OutDegree

These asymmetries in the degree of the nodes are produced by the non-reciprocity of the mentions. For example, 338 references were recorded from <msssi.gob.es> towards <isciii.es>, while 505 were recorded in the opposite direction. The most notable case in point is in the Catalonia subnet, where 15,400 mentions from <prbb.org> towards <imim.es> were recorded, while not a single mention was registered in the opposite direction (data from September 2014; in November 2014 only 77 were obtained).

4.4. Correlation between interlinking and co-authorship data

The Scopus database shows academic output (at least one article published) for a total of 46 institutions. The Catalan Institution for Research and Advanced Studies ranked first (6,753 articles published) followed by the Carlos III Health Institute (5,454) and the Biological Research Centre (4,839). Full details about the academic production of these 46 institutions are provided in the complementary material (Appendix IX). The correlation between the number of articles published by these 46 institutions and the page count of the corresponding institution web domains is lower than expected but statistically significant (0.4; $\alpha=0.01$).

The co-authorship data between each pair of the 46 institutions (454 combinations) was also gathered. The combination with the most activity is that between the Andalusian School of Public Health and Catalan Institute of Oncology (354 papers co-authored). It is also worth noting that within the top 25 institution combinations according to number of co-authored articles, we found thirteen combinations formed by a national and Catalan institution and seven formed by two Catalan institutions. Conversely, only three combinations are formed by two national institutions.

A co-authorship matrix was created (see Appendix X) and compared with the interlinking matrix. The Mantel test between these two matrices shows a significant correlation between the number of co-authored papers and the number of links between the corresponding institution web domains (0.490; $\alpha=0.01$).

These correlations (page count vs academic output; interlinks vs co-authored papers) show a complementary view of institutional relationships as we can observe in Figure 6, where a dispersion graph is provided for each comparison.

Figure 6. Dispersion between page count and academic output data (left) and between interlinking and co-authorship data (right)

Additionally, the co-authorship degree (number of institutions with which a particular institution has collaborated) and interlinking degree (number of web domains to which a particular web domain is linked) were calculated (see Appendix XI). In this case, we obtained a better correlation between these two metrics (0.7; $\alpha=0.01$).

We nonetheless found some important exceptions. For example, the Foundation for the Development of Genomic and Proteomic Research only co-authored articles with 2 other institutions whereas it is hyperlinked with 17 different institution web domains. A similar situation occurred with the Thematic Network of Cooperative Cancer Research (which shares academic publications with 4 institutions though it is hyperlinked with 31). Conversely, the Cardiovascular Research Centre shares publications with 17 institutions while it is hyperlinked with only 7; and the Pere Virgili Health Research Institute publishes with 18 different institutions whereas it is connected with only 7 via the web.

Homogeneous set of research centres

The above correlations were obtained from the interlinking and co-authorship patterns of the 46 institutions that met our study requirements, i.e. between institutions not only with different lines of research but with different institutional goals. In order to observe behaviour amongst a small set of homogeneous institutions, we analysed the institutional relationship patterns between the various Biomedical Research Networking Centres (CIBER), now integrated according to their field of research within the CIBER ISCII consortium.³ Figure 7 maps the co-authorship (left) and interlinking (right) networks for the different centres.

Figure 7. Co-authorship (left) and interlinking (right) networks for Spanish CIBER centres (n=9; Fruchterman Reingold algorithm)

Co-authorship node size: papers published; interlinking node size: page count

The Spearman correlation between the number of articles published jointly and the number of links between each pair of centres is very low ($r=0.3$, $\alpha=0.01$). The correlation in terms of size (number of articles published and page count) is not only equally discrete but even returned a negative value ($r=-0.4$).

5. Discussion

Web data results should be always treated with caution, especially page count and web visibility metrics. Inconsistencies observed and summarised by Wilkinson and Thelwall (2013) apply to any web research which relies on commercial search engines.

Linking root domain seems a more robust and appropriate metric to measure web visibility than total links received, though we must take into account the fact that they actually measure complementary views of visibility (wideness and intensity, respectively). As regards MozRank, while it seems robust as well, this metric suffers from a lack of international coverage that currently prevents it being used for worldwide analyses that include subdomains and subdirectories.

Selective linking through URL mentions (via Google) should be given particular attention. The manual gathering of these data permitted the authors to identify and

classify up to seven types of errors associated with this metric, some of them previously detected in the literature (Thelwall and Sud 2011; Ortega et al. 2014; Sud and Thelwall 2014). Notwithstanding, we have detailed them below to give a better understanding of our study:

a) High variability

For example, the difference in the number of mentions from <imim.es> towards <prbb.org> between the September (0 entries) and November (77) data collections is large. This may be the result of collaboration between the institutions, the follow-up of an ad hoc event, or improvements or changes made to the coverage of the search engine.

b) The detection of email mentions

For example, for the query <<"cnic.es" site:iccc.cat>, Google detects the mention "xxx@cnic.es". Although this is not a false positive, since it is a mention as such, the nature of an email is different to that of a hyperlink to <cnic.es>, which must be taken into account.

c) The search string of the target does not appear in the source

For example, the query <<"prbb.org" site:icrea.cat> produced 7 results. After a manual check, we found that the string "irbbarcelona.org" did not appear in any of the results, which means that, in this case, false positives were given. This is because the source (<idibgi.org>) is a dynamic web resource and its contents vary periodically, so the string perhaps appeared at the time that the crawler indexed the page, but subsequently disappeared, as the index had not been updated at the time the query was made.

d) Errors in string detection

Despite the use of quotation marks to delimit the strings to be detected, the search engine did not properly identify spaces, dots and other symbols and diacritics, thus returning false positives. This was a particularly problematic issue for domains with the <.es> suffix, since "es" is Spanish for "is" (third-person singular form of the verb "to be"). Hence phrases like "Idibell *es* un centro (*is* a centre)...", "PREDIMED *es* el acrónimo (*is* an acronym)...", or "El Instituto de Salud Carlos III (ISCIII) *es* un (*is* a)...", are detected as <idibell.es>, <predimed.es> or <isciii.es> respectively. These domains consequently obtained slightly inflated results.

e) Different language versions

Those websites with versions in different languages (especially in regions with co-official regional languages) generated a result for each of the languages used. For example, the query <<"recava.com" site:idibaps.org> generated 3 results, which are

actually the same, but repeated in the three language versions of the website <idibaps.org> (Spanish, Catalan and English).

f) Targets in second- and third-level subdomains

Occasionally, when measuring a particular subdomain, we are actually measuring a subdomain hosted within it. For example, all the results returned for the query <“san.gva.es” site:isglobal.org> are actually hosted on <sp.san.gva.es>, which is an entity for which data was gathered separately.

g) Target and source hosted by the same web domain

When considering distinct units hosted on different subdomain levels in the same web domain, the system is unable to calculate mentions among them with precision.

For example, when counting mentions to <bioef.org>, the system returned all strings with these terms, incorporating all existing <xxx.bioef.org> subdomains, e.g. <oiker.bioef.org>. Therefore, it is impossible to correctly determine the number of mentions that are solely directed towards the general unit and, more specifically, those from the lower unit to the general unit (and vice versa). This effect particularly affects the performance of the Ministry of Health, Social Services and Equality (<msssi.gob.es>), which assumes a more central role in the area of public health: it contains a large amount of services, centres and institutions, and also includes navigation links between all of them.

These limitations especially affect the use of these indicators to evaluate performance (the exact number of references to a URL). However, if used for relational purposes (to determine whether the relationship between “a” and “b” is greater than that between “a” and “c”, for example), these limitations are minimised since all URLs are subject to the same error, and thus the error is statistically dispersed. Therefore we consider that the method efficiently detects combinations in which the intensity of the relationship is high despite potential specific errors. Moreover, even assuming a minimal rate of error, given the general low URL mentions obtained, the effect on network indicators is practically imperceptible, as they remain unaffected by changes in the intensity of the relationship.

6. Conclusions

Institutional relationships between Spanish public health entities have been identified both through link analysis and co-authorship methods. We found a discrete but statistically significant correlation between the number of selective links across a pair of entities and the number of academic articles published jointly by them ($r=0.5$).

This value suggests a complementary, though not alternative, view of each method in the identification of institutional relationships. The existence of institutions with scarce academic collaboration but an intense web relationship (reflecting a relationship based on informative and web navigation) and, conversely, institutions with an intense academic relationship but insufficient web connection (reflecting the existence of inadequate policies for creating web content), reinforces this issue.

Notwithstanding, if we perform the comparison by means of the degree indicator (interlinking and co-authorship degree) the correlation clearly improves ($r=0.7$). This fact suggests that hyperlinks are better suited for identifying entities that, academically, collaborate intensively with many others instead of measuring with precision the intensity of a particular relationship.

Moreover, both interlinking and co-authorship methods detect the existence of strength subnets of geographically distributed nodes (especially the Catalan entities) as well as their high connectivity with the main national network nodes (subnet of nodes distributed according to dependence on national government, in this case Spain). Each method is capable additionally of identifying different connection purposes within each subnet. In this specific case, whereas the Catalan network is as well connected through hyperlinks as it is through academic publications, the national network is better connected through hyperlinks. This fact may be due to the role adopted by hyperlinks at a national level, mainly devoted to navigational purposes.

In any case, the general network of web relationships between the analysed Spanish public health entities has very little cohesion, is highly distributed, inadequately connected, and the relationships between nodes are not particularly intense. The dissociation between page count and web visibility metrics provided by the web impact analysis (and reflected in the moderate correlation obtained between them) may suggest that the creation of web content is not influencing the attraction of external hyperlinks, which may explain the general low cohesion of the network.

The widespread use of unsuitable URL syntaxes (subdirectories and unfriendly forms of subdirectory and combined URL syntaxes) may also have an influence. These URLs, despite functioning on a technical level, do not facilitate the generation of links by users or correct quantification by search engines.

The web relationships between a parent institution and its subsidiaries generate significant distortion because of the large number of existing navigation links. This means that link analysis reinforces institutional relationships based on informative communication and access to content. However, this is the only way to approach a multilevel analysis using webometric indicators (since subdomains reflect relationships of institutional affiliation). Suitable filtering of these links could provide a greater granularity between these parent/subsidiary relationships, although this task is tedious and difficult to automate. Therefore, relationships between institutions with different web domains will always be more methodologically precise.

Finally, the analysis of small samples of specialised research centres (in this case, the CIBER centres) has shown that institutional relationships detected by each method may differ even among thematically (biomedical research) and functionally related centres (all linked to the Spanish National Health System). There may be web relationships (functional, administrative, communicative, etc.) without there being a scholarly relationship (the centres do not publish articles jointly). However, the existence of scholarly relationships that lack correspondence on the web is evidence of gaps in the creation and adequate dissemination of content in the corresponding centres. Hence the combination of the two methods of applied analysis not only serves to identify a wider

range of institutional relationships but also to highlight gaps in the generation of content that is indicative of existing relationships.

Greater planning of web policies for the creation and dissemination of the web content of entities related to the Spanish healthcare sector would result in an improvement, not only in web impact indicators, but also in obtaining a richer and more intense network of relationships. This would more accurately reflect the potential relationships between the various entities (including academic relationships) and minimise the limitations of URL mentions. Finally, social metrics reflecting institutional relationships – and not based on hyperlinks – may provide complementary insights; this is an aspect that will therefore be considered in future studies.

Endnotes

1. <https://moz.com/researchtools/ose/>
2. Some institutions (such as the Valencia Institute for Health Research-INCLIVA, the La Fe Health Research Institute-IISLAFE or the Institute of Biomedical Research of Barcelona-IIBB) were excluded from the sample because during the process of data collection they were not hyperlinked by the source websites. In any case, in Appendix III, hyperlink, web impact and correlation data (co-authorship vs interlinking) for these three institutions (as of November 2015) are shown. As can be seen, institutional relationships between these institutions (both in co-authorship and, primarily, interlinking) are weak, so their exclusion does not affect the results presented in this paper or its main conclusions. Notwithstanding, these data are shown for interested readers.
3. <http://www.ciberisciii.es>

References

- Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for information science and technology*, 57(10), 1296-1302.
- Almind, T. C., & Ingwersen, P. (1998). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404-426.
- Barabasi, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439), 509-512.
- Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management*, 41(4), 973-986.
- Barnett, George A., & Park, Han W. (2014). Examining the international internet using multiple measures: new methods for measuring the communication base of globalized cyberspace. *Quality & Quantity*, 48(1), 563-575.
- Eurostat. (2011). *Regions in the European Union. Nomenclature of territorial units for statistics*. NUTS 2010/EU-27.
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-11-011/EN/KS-RA-11-011-EN.PDF (accessed 16 August 2015).
- García-Lacalle, J., Pina, V., & Royo, S. (2011). The unpromising quality and evolution of Spanish public hospital web sites. *Online information review*, 35(1), 86-112.

PREPRINT OF ARTICLE ACCEPTED FOR PUBLICATION IN
SCIENTOMETRICS

DOI: 10.1007/s11192-016-1839-z

García-Santiago, L., & Moya-Anegón, F. (2009). Using co-outlinks to mine heterogeneous networks. *Scientometrics*, 79(3), 681-702.

González-Bailón, S. (2009). Opening the black box of link formation: social factors underlying the structure of the web. *Social Networks*, 31(2009), 271-280.

Heimeriks, G., Hörlesberger, M., & Van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2), 391-413.

Heimeriks, G., & Van den Besselaar, P. (2006). Analyzing hyperlinks networks: The meaning of hyperlink based indicators of knowledge production. *Cybermetrics*, 10(1), <http://cybermetrics.cindoc.csic.es/articles/v10i1p1.pdf> (accessed 16 August 2015).

Holmberg, K. (2010). Co-inlinking to a municipal Web space: A webometric and content analysis. *Scientometrics*, 83(3), 851-862.

Holmberg, K., & Thelwall, M. (2009). Local government web sites in Finland: a geographic and webometric analysis. *Scientometrics*, 79(1), 157-169.

Khan, G. F., & Park, H. W. (2011). Measuring the triple helix on the web: longitudinal trends in the university-industry-government relationship in Korea. *Journal of the American Society for Information Science and Technology*, 62(12), 2443-2455.

Lang, P. B., Gouveia, F. C., & Leta, J. (2014). Health research networks on the web: an analysis of the Brazilian presence. *Cadernos de Saúde Pública*, 30(2), 369-378.

Leydesdorff, L., & Curran, M. (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy. *Cybermetrics*, 4(1). <http://www.cybermetrics.info/articles/v4i1p2.pdf> (accessed 16 August 2015).

Méndez-Vásquez, R. I., Suñen-Pinyol, E., Cervelló, R., & Camí, J. (2008). Mapa bibliométrico de España 1996-2004: biomedicina y ciencias de la salud. *Medicina clínica*, 130(7), 246-253.

Méndez-Vásquez, R. I., Suñen-Pinyol, E., & Rovira, L. (2012). Caracterización bibliométrica de la investigación biomédica española, WOS 1997-2011. <http://bac.fundacionrecerca.cat/mb11> (accessed 16 August 2015).

Ministerio de Sanidad, Servicios Sociales e Igualdad. (2012), *Sistema Nacional de Salud. España 2012*. http://www.msssi.gob.es/organizacion/sns/docs/sns2012/SNS012__Espanol.pdf (accessed 16 August 2015).

Orduna Malea, E., Ortega, J. L., & Aguillo, I. F. (2014). Influence of language and file type on the web visibility of top European universities. *Aslib Proceedings*. 66(1), 96-116.

Orduna-Malea, E., & Aguillo, I. F. (2014). *Cibermetría. Midiendo el espacio red*. Barcelona: UOC Publishing.

Orduna-Malea, E., & Aytac, S. (2015). Revealing the online network between university and industry: the case of Turkey. *Scientometrics (First online)*.

Orduna-Malea, E., & Ontalba-Ruipérez, J. A. (2013). Proposal for a multilevel university cybermetric analysis model. *Scientometrics*, 95(3), 863-884.

PREPRINT OF ARTICLE ACCEPTED FOR PUBLICATION IN
SCIENTOMETRICS

DOI: 10.1007/s11192-016-1839-z

Orduna-Malea, E., Delgado López-Cózar, E., Serrano-Cobos, J., & Romero, N. L. (2015). Disclosing the network structure of private companies on the web: the case of Spanish IBEX 35 share index. *Online Information Review*, 39(3), 360-382.

Orduna-Malea, E., Torres-Salinas, D., & Delgado López-Cózar, E. (2015). Hyperlinks embedded in twitter as a proxy for total external in-links to international university websites. *Journal of the Association for Information Science and Technology*, 66(7), 1447-1462.

Ortega, J. L. (2007). *Visualización de la Web universitaria Europea: análisis cuantitativo de enlaces a través de técnicas cibernéticas*, Madrid: Universidad Carlos III de Madrid.

Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management*, 45(2), 272-279.

Ortega, J. L., Orduna-Malea, E., & Aguillo, I. F. (2014). Are web mentions accurate substitutes for inlinks for Spanish universities?. *Online Information Review*, 38(1), pp. 59-77.

Park, H. W., & Thelwall, M. (2003). Hyperlink analyses of the World Wide Web: A review. *Journal of Computer-Mediated Communication*, 8(4).
<http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2003.tb00223.x> (accessed 16 August 2015).

Park, H.W. (2011). How do social scientists use link data from search engines to understand Internet-based political and electoral communication?. *Quality & Quantity*, 46(2), 679-693.

Romero-Frías, E., & Vaughan, L. (2010a). Patterns of web linking to heterogeneous groups of companies: The case of stock exchange indexes. *Aslib Proceedings*, 62(2), 144-164.

Romero-Frías, E., & Vaughan, L. (2010b). European political trends viewed through patterns of Web linking. *Journal of the American Society for Information Science and Technology*, 61(10), 2109-2121.

Seeber, M., Lepori, B., Lomi, A., Aguillo, I. F., & Barberio, V. (2012). Factors affecting web links between European higher education institutions. *Journal of informetrics*, 6(3), 435-447.

Stuart, D., & Thelwall, M.(2006). Investigating triple helix relationships using URL citations: A case study of the UK West Midlands automobile industry. *Research evaluation*, 15(2), 97-106.

Sud, P., & Thelwall, M. (2014). Linked title mentions: a new automated link search candidate. *Scientometrics*, 101(3), 1831-1849.

Thelwall, M. (2001). Extracting macroscopic information from web links. *Journal of the American Society for Information Science and Technology*, 52(13), 1157-1168.

Thelwall, M. (2002). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation*, 58(5), 563-574.

Thelwall, M. (2004). *Link Analysis: An Information Science Approach*. San Diego: Elsevier.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, 57(1), 60-68.

**PREPRINT OF ARTICLE ACCEPTED FOR PUBLICATION IN
SCIENTOMETRICS**

DOI: 10.1007/s11192-016-1839-z

Thelwall, M. (2009). *Introduction to webometrics: quantitative web research for the social sciences*. San Rafael, CA: Morgan & Claypool Publishers.

Thelwall, M., & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organisations. *Journal of the American Society for Information Science and Technology*, 62(8), 1488-1497.

Thelwall, M., & Tang, R. (2003). Disciplinary and linguistic considerations for academic web linking: an exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics*, 58(1), 155-181.

Thelwall, M., Tang, R., & Price, L. (2003). Linguistic patterns of Academic web use in Western Europe. *Scientometrics*, 56(3), 417-432.

Vaughan, L. (2006). Visualizing linguistic and cultural differences using web co-link data. *Journal of the American Society for Information Science and Technology*, 57(9), 1178-1193.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the web: what are the key inducers of links to journal web sites?. *Journal of the American Society for Information Science and Technology*, 54(1), 29-38.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information processing & management*, 40(4), 693-707.

Vaughan, L., & Wu, G. (2004). Links to commercial websites as a source of business information. *Scientometrics*, 60(3), 487-496.

Vaughan, L., & You, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, 68(3), 611-628.

Weber, M. S., & Monge, P. (2011). The flow of digital news in a network of sources, authorities, and hubs. *Journal of Communication*, 61(6), 1062-1081.

Wilkinson, D., & Thelwall, M. (2013). Search markets and search results: The case of *Bing*. *Library & Information Science Research*, 35(4), 318-325.

Wilkinson, D., Harries, G., Thelwall, M., & Price, L. (2003). Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of information science*, 29(1), 49-56.

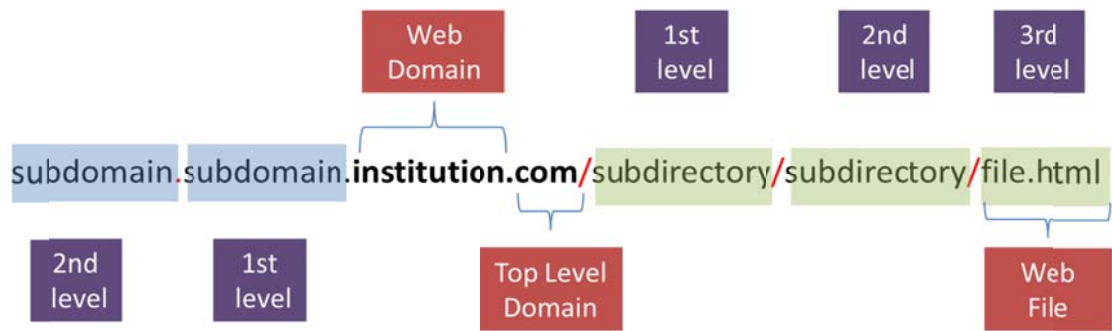


Figure 1. URL syntax nomenclature

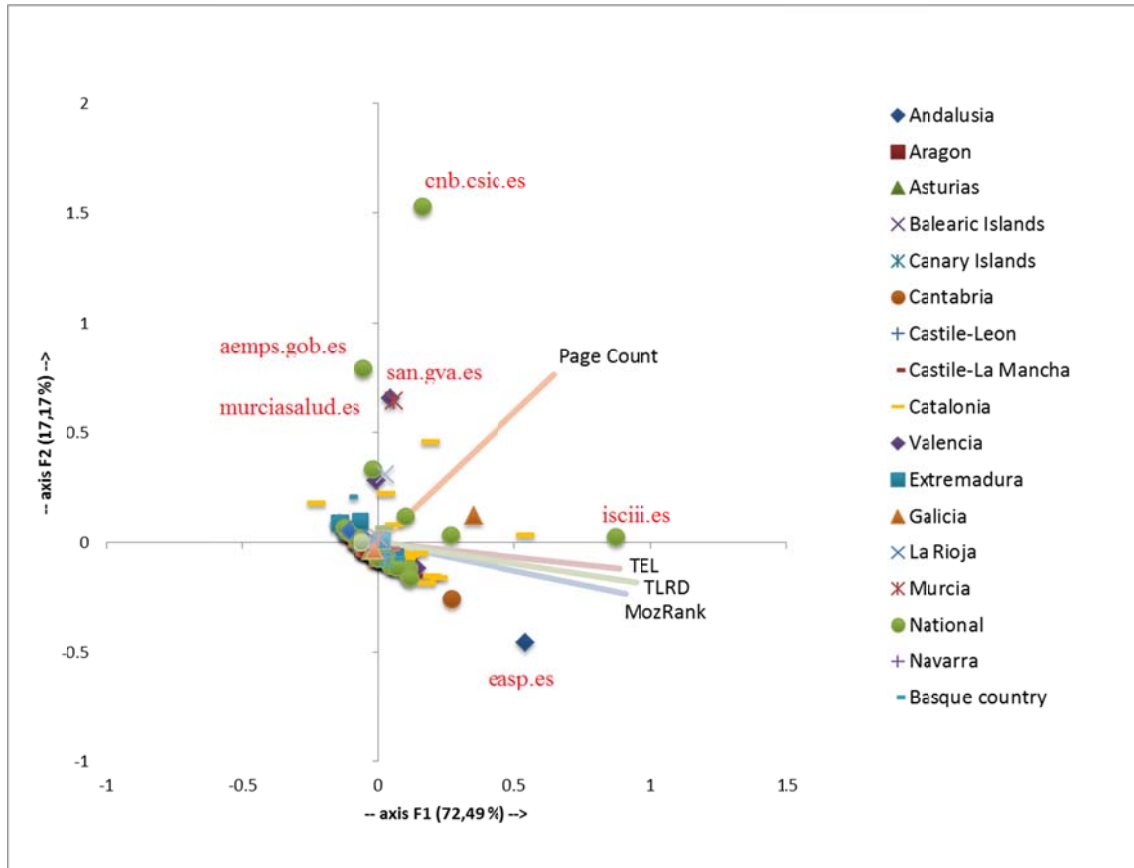
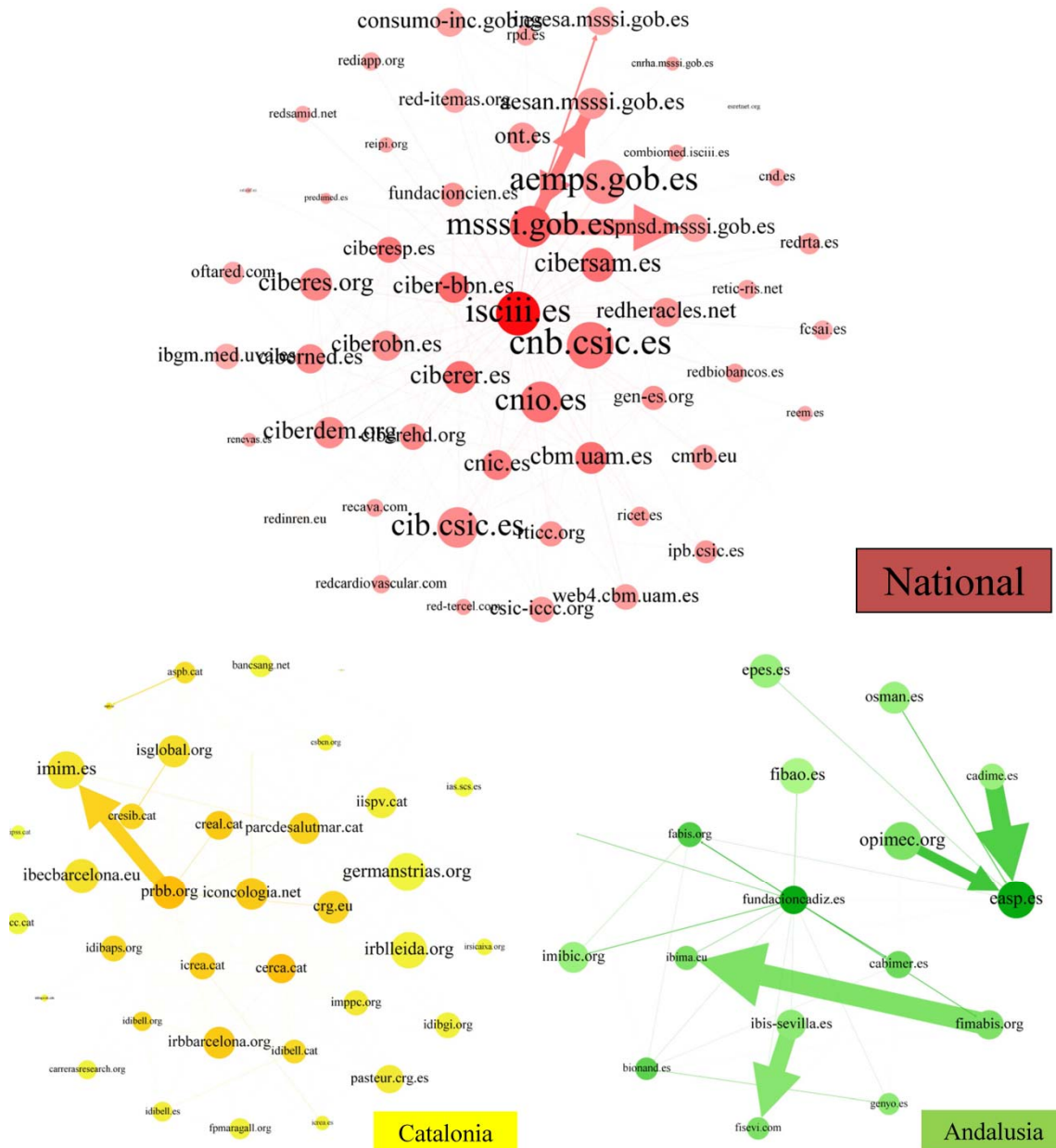


Figure 2. Principal Component Analysis (PCA) of units according to region



Metric	Total (n= 175)	National (n=55)	Catalonia (n= 38)	Andalusia (n= 18)
Average Degree	10.046	6.8	5.47	2.11
Diameter	6	4	5	4
Density	0.058	0.126	0.148	0.124
Average Clustering Coeff.	0.411	0.527	0.332	0.195
Average Path Length	2.356	2.077	2.256	1.916

Figure 5. Interconnection of the national and Catalan subnets

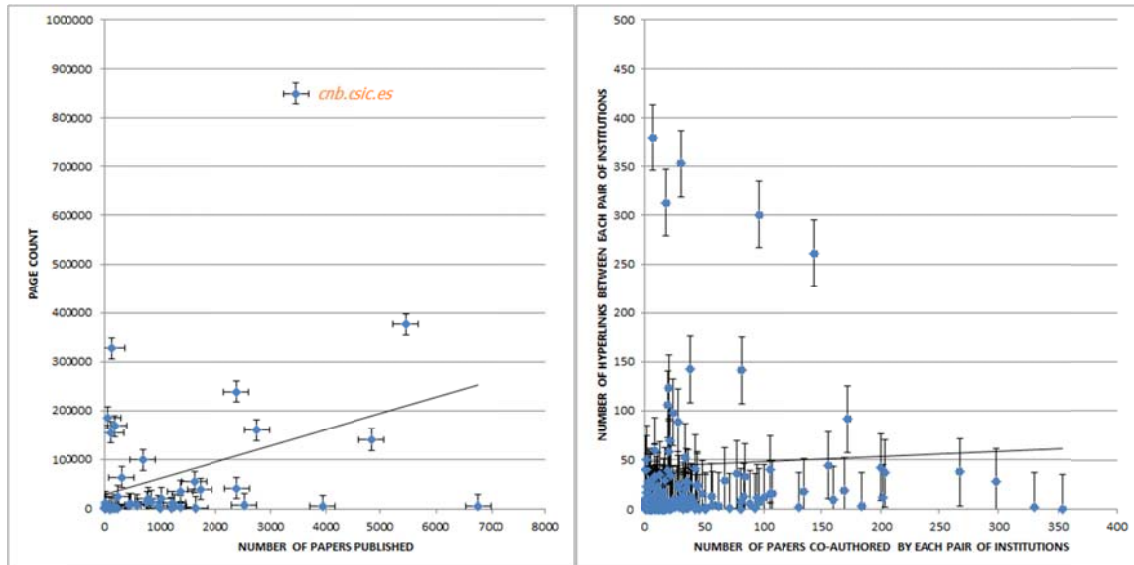


Figure 6. Dispersion between page count and academic output data (left) and between interlinking and co-authorship data (right)

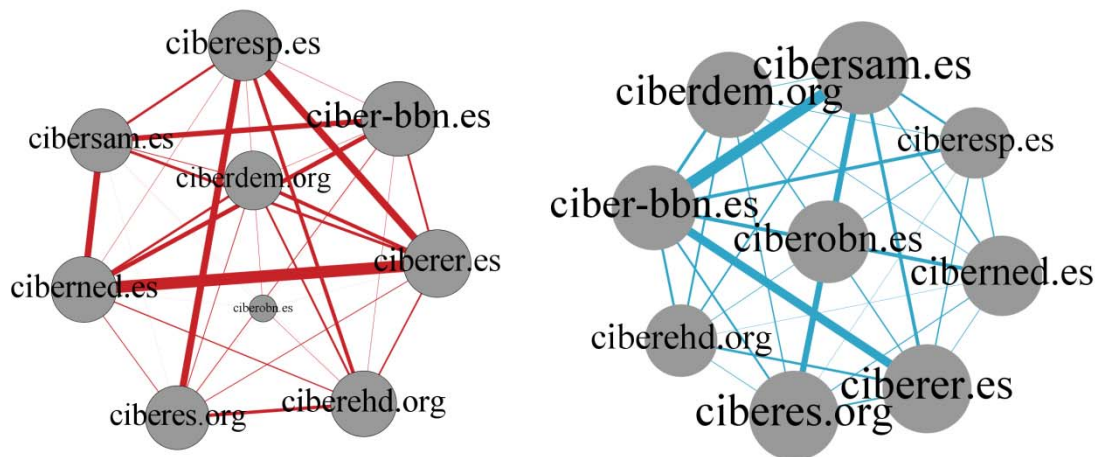


Figure 7. Co-authorship (left) and interlinking (right) networks for Spanish CIBER centres (n=9; Fruchterman Reingold algorithm)

Co-authorship node size: papers published; interlinking node size: page count