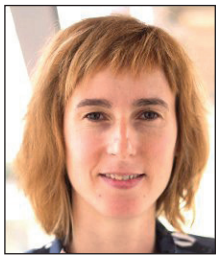
# SEMI-AUTOMATIC GENERATION OF A CORPUS OF *WIKIPEDIA* ARTICLES ON SCIENCE AND TECHNOLOGY

## Generación semi-automática de un corpus de artículos de Wikipedia sobre ciencia y tecnología

**Julià Minguillón, Maura Lerga, Eduard Aibar, Josep Lladós-Masllorens and Antoni Meseguer-Artola**

**Julià Minguillón** received his PhD degree from the *Universitat Autònoma de Barcelona* (*UAB*) in 2002. In January 2001 he joined the *Universitat Oberta de Catalunya* (*UOC*) faculty in the *Computer Science, Multimedia and Telecommunication Studies Department*. He is member of the *UOC*'s *Laika* research group. His research interests include the uses of open educational resources, digital repositories, and social tools for teaching and learning in virtual learning environments, educational data mining, learning analytics, and data visualization.
*http://orcid.org/0000-0002-0080-846X*

*Universitat Oberta de Catalunya*
Rambla Poblenou, 156. 08018 Barcelona, Spain
*jminguillona@uoc.edu*



**Maura Lerga**, sociologist, holds a master's on Urban Anthropology. She is an experienced research assistant with a demonstrated history of working in the higher education field. She is skilled in sociology, data analysis, academic writing, strategic planning, and project management with a strong background as a social consultant. She is currently working at the *International Center* of the *Universitat Rovira i Virgili* (Spain), where she is in charge of four *Erasmus+* projects, as project manager. She is a former member of the *Open Science and Innovation Research Group* (Universitat Oberta de Catalunya), where she was studying the scientific controversies in *Wikipedia*.
*http://orcid.org/0000-0002-3516-0931*

*Universitat Oberta de Catalunya*
Avda. Tibidabo, 39-43. 08035 Barcelona, Spain
*mlergaf@uoc.edu*

**Eduard Aibar** is an associate professor of Science and Technology Studies at the *Arts & Humanities Department*, *Universitat Oberta de Catalunya* (*UOC*). He has been the director of the *Internet Interdisciplinary Institute* and vice-president for research at *UOC*. He leads a research group on open science and innovation. His research has focused on the interaction between scientific and technological development and organizational and social change in areas such as eGovernment, town planning, and the Internet. He is currently leading a research project on the relationship between science and peer production.
*http://orcid.org/0000-0002-1727-1523*

*Universitat Oberta de Catalunya*
Avda. Tibidabo, 39-43. 08035 Barcelona, Spain
*eaibar@uoc.edu*

**Josep Lladós-Masllorens** is an associate professor of the *Business and Economics Department* at the *Universitat Oberta de Catalunya* (*UOC*), Barcelona (Spain). In 1998, he received a PhD in Business and Economics Science from the *Universitat de Barcelona*. He is a member of the *Digital Business Research Group* (*DigiBiz*), a consolidated research team recognized by the *Generalitat of Catalonia*. His research interests include economic geography, innovation systems, global value chains, and sharing economy; he is the author of several scientific papers, reports, and book chapters on these topics.
*http://orcid.org/0000-0002-6236-5166*

*Universitat Oberta de Catalunya*
Avda. Tibidabo, 39-43. 08035 Barcelona, Spain
*jlladosm@uoc.edu*

Manuscript received on 25-05-2017
Accepted on 19-06-2017

**Antoni Meseguer-Artola** has a doctoral degree in Economics and Business Sciences, a master's degree in Economic Analysis from the *UAB,* and a bachelor's degree in Mathematics from the *UB*. He is an associate professor of Quantitative Methods for Economics and Business at the *UOC* and a member of the *DigiBiz* research group. He is author of several articles and conferences on digital marketing, consumer behavior, game theory, and e-learning. He has a merit in research. He has also authored several teaching books in the field of mathematics and statistics. At the *UOC*, he has been director of the *Bachelor's Degree Program in Work Sciences* (2001-2006) and dean of the *Faculty of Economics and Business* (2006-2010).
*http://orcid.org/0000-0002-7817-3695*

*Universitat Oberta de Catalunya*
Avda. Tibidabo, 39-43. 08035 Barcelona, Spain
*ameseguer@uoc.edu*

**Abstract**

Despite the huge amount of scientific and technological content available on the World Wide Web, most of it is closed behind paywalls, as with academic journals, or almost invisible, as with institutional repositories. *Wikipedia* can act as a chain-transfer agent, providing people with an accessible, organized structure containing both understandable content and links to original sources. In *Wikipedia*, categories are collaboratively created and thus become a folksonomy rather than a true taxonomy. Consequently, categories are not a reliable tool to identify topics' organization. In this paper we describe a semi-automatic method, based on random walks, for determining a subset of pages containing scientific and technological content in the Spanish *Wikipedia*. Using the *Unesco* taxonomy, we determined the underlying graph structure of our corpus and detected clusters of pages strongly linked, establishing relationships between knowledge domains. Finally, we present the distribution of *Wikipedia* articles according to the *Unesco* taxonomy and the resulting map of scientific and technological content.

**Keywords**

*Wikipedia*; Science and technology; Corpus; *Infomap*; Community detection; *Unesco* taxonomy.

**Resumen**

A pesar de la gran cantidad de contenido científico y tecnológico disponible en la World Wide Web, su mayoría se encuentra encerrado tras sistemas de pago, como las revistas académicas, o es casi invisible, como los repositorios institucionales. *Wikipedia* puede actuar como un agente de transferencia, proporcionando una estructura organizada y accesible conteniendo tanto contenidos como enlaces a las fuentes originales. En *Wikipedia* las categorías se han cread colaborativamente y por lo tanto son más una folksonomía que una verdadera taxonomía. Consecuentemente, las categorías no son una herramienta válida para identificar la organización de los contenidos. En este artículo se describe un método semi-automático, basado en paseos aleatorios, para determinar un subconjunto de páginas con contenido científico y tecnológico de la *Wikipedia* española. Usando la taxonomía *Unesco*, se determina la estructura subyacente del grafo del corpus y se detectan grupos de páginas fuertemente enlazadas, estableciendo las relaciones entre las áreas de conocimiento. Finalmente, se presenta la distribución de artículos de *Wikipedia* de acuerdo con la taxonomía *Unesco* y el mapa resultante de contenido científico y tecnológico.

**Palabras clave**

*Wikipedia*; Ciencia y tecnología; Corpus; *Infomap*; Detección de comunidades; Taxonomía *Unesco*.

**Minguillón, Julià**; **Lerga, Maura**; **Aibar, Eduard**; **Lladós-Masllorens, Josep**; **Meseguer-Artola, Antoni** (2017). "Semi-automatic generation of a corpus of *Wikipedia* articles on science and technology". *El profesional de la información*, v. 26, n. 5, pp. 995-1004.

*https://doi.org/10.3145/epi.2017.sep.20*

## 1. Introduction

*Wikipedia* has become one of the most popular websites in the World Wide Web. Due to its highly linked hypertext structure, it is a resource that is accessible through search engines, including *Google* (**Ermann**; **Frahm**; **Shepelyansky**, 2015). It is ranked in seventh position in the Internet rankings of *Alexa*, for instance. In fact, almost any search in Google returns one or more *Wikipedia* pages on the first results page, and often in first position, thus promoting its usage. Therefore, *Wikipedia* has rapidly become the most used reference source for any user searching for information about any topic (**Ponzetto**; **Strube**, 2007). On the other hand, *Wikipedia* has also become an excellent resource for research purposes (**Medelyan** *et al.*, 2009), including both content and semantic analysis (**Ryu**; **Jang**; **Kim**, 2014), as it provides a huge source of organized human knowledge.

> "*Wikipedia* has become the main platform for the public communication of science"

A less well-known fact about *Wikipedia* is that it has also become the main platform for the public communication of science. Recent studies on communication and public perception of science prove that the number of people who rely on *Wikipedia* as a source of information about science and technology has rapidly increased in recent years and has surpassed all other media (**Brossard**; **Scheufele**, 2013; **Snyder**, 2013). Among students, *Wikipedia* is the main information source (**Kim**; **Sin**; **Yoo-Lee**, 2014). Even faculty and scientists themselves have become frequent users of *Wikipedia* in searching for scientific and academic issues (**Aibar** *et al.*, 2015; **Snyder**, 2013).

Considering its increasing influence and use, the analysis of the scientific and technological content of *Wikipedia* in terms of quality, completeness, reliability and bias, to mention just a few aspects, is of great interest for many fields including science studies, public understanding of science, information management, etc. But for that analysis a previous step is needed: the identification of the corpus of *Wikipedia* articles dealing with science and technology issues. This is indeed the problem we have addressed in this paper and, for the reasons we will now show, it proves to be a non-trivial and hard task.

> The Spanish language edition of *Wikipedia* has more than 1.2 million articles and is the 9th largest, with more than 15,000 active users

Regarding its content, *Wikipedia* includes more than five million articles in its English edition, and there are versions in 282 languages (as of June 2016). More than 28 million people have contributed to *Wikipedia*, and more than 116,000 can be considered active users, meaning they have edited content during the last month. It is, therefore, the biggest multilingual collaborative effort ever made (**Samoilenko** *et al.*, 2016). In the particular case of the Spanish language edition of *Wikipedia*, it has more than 1.2 million articles and is the 9th largest, with more than 15,000 active users in 2017. According to the *Instituto Cervantes*, there are almost 468 million people that have Spanish as their mother tongue, thus its importance as a channel for accessing scientific and technological knowledge. Furthermore, most research on *Wikipedia* focuses exclusively on the English version and studies are needed in other editions (**Mesgari** *et al.*, 2015). Along this line, **Figuerola**, **Groves,** and **Quintanilla** (2015) have analyzed the Spanish *Wikipedia* as an educational tool for scientific and technological content, using a completely manual approach in a data dump dated November 2013. Our paper extends the work of **Figuerola**, **Groves,** and **Quintanilla** (2015) by using pre-existing taxonomies for identifying scientific and technological content in a more recent dump dated December 2014.

*Wikipedia* uses categories to classify articles in a huge pseudo-taxonomy, which has also been collaboratively created (**Muchnik** *et al.*, 2007; **Thornton**; **McDonald**, 2012; **Kaptein**; **Kamps**, 2013). Although the quality of content in *Wikipedia* pages is comparable to other encyclopedias (**Lih**, 2004), this cannot be said about the taxonomy of categories used to describe its content (**Silva** *et al.*, 2011; **Salah** *et al.*, 2012; **Hejazy**; **El-Beltagy**, 2013). Several authors have stated problems with the *Wikipedia* category system. Among them, **Capocci**, **Rao,** and **Caldarelli** (2007) analyzed the structure imposed by categories, showing that it did not reproduce the natural division of articles according to their links. **Halavais** and **Lackaff** (2008) showed that *Wikipedia* coverage depended on its users' interests, as well the category system used to organize them. **Holloway**, **Božicevic,** and **Börner** (2007) described some of the known inconsistencies of categories, such as lack of hierarchical structure or the presence of cycles (**Muchnik** *et al.*, 2007). As stated in (**Suchecki** *et al.*, 2012), taxonomies have been more stable in the bottom (that is, the category terms in *Wikipedia* pages do not change over time) than in the top-level terms, which may be reorganized occasionally. Furthermore, the taxonomy is not consistent across languages, making the creation of multilingual collections for a given concept or domain very difficult. As several authors have pointed out, this is mainly because *Wikipedia* categories are more a folksonomy rather than a true taxonomy (**Jiménez-Pelayo**, 2009; **Hejazy**; **El-Beltagy**, 2013). Nevertheless, and despite these facts, *Wikipedia* categories have been extensively used to extract ontological knowledge from *Wikipedia* (**Kaptein**; **Kamps**, 2013; **Ryu**; **Jang**; **Kim**, 2014; **Jiang** *et al.*, 2015), showing its importance as a huge ground-truth data set for building knowledge-based systems.

However, some precautions should be taken when working with *Wikipedia* categories. For instance, the Spanish *Wikipedia* contains a top-level page for categories, including a category for 'Science' (linking to 13 subcategories and 15 pages) and another for 'Technology' (30 subcategories and 187 pages). In the case of 'Science', the 13 subcategories point to 636 subcategories, including non-scientific content (i.e. 'Pseudoscience'). In the case of 'Technology', the 30 subcategories point to 367 subcategories. Furthermore, there are other possible entry points providing access to some of these subcategories, such as 'Academic disciplines', so they cannot be considered to be orthogonal or mutually exclusive. As shown by **Suchecki** *et al.* (2012) and **Hejazy,** and **El-Beltagy** (2013), some cleansing is needed in order to obtain a valid hierarchized taxonomy.

> The *Wikipedia* categories are more a folksonomy rather than a true taxonomy

Therefore, we decided not to rely only on categories to analyze the *Wikipedia* content structure in order to determine scientific and technological content. It would be interesting to identify all the *Wikipedia* pages related to a given concept, without having to manually review the taxonomy of *Wikipedia* categories and deciding whether each category and the pages it points to are truly related to that concept or not.

In this paper we describe a semi-automatized method for creating a corpus of *Wikipedia* articles devoted to scientific and technological content. As we cannot rely on *Wikipedia*

taxonomies, we will combine the use of community detection techniques based on random walks, extracting information from the internal structure of *Wikipedia* links between pages following the approach described by **Figuerola**, **Groves,** and **Quintanilla** (2015), but including the use of existing taxonomies accepted as a *de facto* standard for describing academic content. The research questions addressed in this paper are:

Q1: Can a significant subset of *Wikipedia* pages related to scientific and technological content be semi-automatically identified without using the existing category structure?

Q2: Does such subset contain comprehensive coverage of scientific and technological content or is it biased towards some specific fields?

## 2. Methodology

Due to its encyclopedic and hyperlinked nature, most *Wikipedia* pages link to other pages that can be regarded as being devoted to the same topic, as they explain or introduce related concepts. For instance, 'oxygen' links to 'chemical element', 'atomic number', and 'periodic table', among others. These pages point back to 'oxygen' and other chemical elements as well, creating a dense structure of links between them. Therefore, any random walk starting in 'oxygen' will probably visit several *Wikipedia* pages about other chemical elements, discovering an underlying community structure (**Pons**; **Latapy**, 2005), which can be used to automatically categorize *Wikipedia* pages. Notice that, throughout this paper, we use "community" as the name for any subset of strongly connected *Wikipedia* pages that are related to the same topic.

> In this paper we describe a semi-automatized method for creating a corpus of *Wikipedia* articles devoted to scientific and technological content

### 2.1. Building the graph

In our case, we are only interested in detecting the underlying graph structure between pages and links, so we will use only the XML dump file containing *Wikipedia* pages. We have analyzed a dump corresponding to the Spanish version of *Wikipedia*, taken on the 17th of December, 2014. The XML file occupies more than nine GB and it contains more than three million pages and 60 million links to other *Wikipedia* pages.

The first step is to recreate the hyperlinked structure of web pages, which allows *Wikipedia* users to jump from one page to another. Pages include *Wikipedia* articles but also categories, lists, user pages, discussion and so on, which cannot be considered real content for the purposes of our analysis. Therefore, if we want to generate a graph containing only pure content, i.e. articles, we need to remove these special pages. On the other hand, we will include annex pages as they are usually lists of *Wikipedia* pages related to the same concept. This can easily be done as these special pages usually use a particular prefix that identifies them. For instance, categories are *Wikipedia* pages that always have 'Category:' as a prefix before the category page ('Categoría:' in Spanish). Analogously, we also remove pages containing files, images, templates, and other pages related to *Wikipedia* (i.e. documentation). Disambiguation pages can also be removed as their names always finish with '_(desambiguación)'. Then we resolve all issues related to redirections, including acronyms, different pages pointing to the same one, spaces, and capital letters, as well as taking care of special characters (like 'Ñ' in Spanish) which may have been removed or replaced in order to ensure proper URL functioning. Finally, as two pages can be linked by more than one jump (i.e. the linked page can appear several times in the starting page), we remove all duplicated links but one, maintaining this relationship. We also remove edges linking a page to itself, as they have no meaning for the purposes of our analysis.

> *Wikipedia* can be seen as a scale-free network

The resulting graph has 1,143,024 nodes (that is, *Wikipedia* pages) and 25,469,978 edges (that is, links between nodes). As stated, we did not include 225,380 category pages. Due to its nature, *Wikipedia* can be seen as a scale-free network (**Barabási**; **Albert**, 1999), where the distribution of node out-degrees (i.e. number of links) emerges automatically from a stochastic growth model in which new nodes are added continuously and attach themselves preferentially to existing nodes, with probability proportional to the degree of the target node--so richly connected nodes get richer (**Strogatz**, 2001), becoming hubs to hundreds or even thousands of pages (like annex pages).

### 2.2. Analyzing the graph structure

Among the several algorithms that can be used to detect communities, we have followed the same approach described in **Bohlin** *et al.* (2014), based in the *Infomap* algorithm (**Rosvall**; **Bergstrom**, 2008). *Infomap* clusters tightly interconnected nodes into modules (two-level clustering) or the optimal number of nested modules (multi-level clustering). It is known to be a fast method that outperforms other algorithms for community detection while providing a high modularity index, which is very suitable for large graphs (**Emmons** *et al*., 2016).

For each node in G (i.e., a page), we obtain a sequence of nested communities and subcommunities that determine the category for that page. The communities form a hierarchical tree, T, with non-overlapping nodes, so each page belongs to a unique community, represented by a leaf on the tree. Obviously, each page also belongs to all other internal nodes that connect the tree root (that contains all communities) with the leaf. Communities are coded $C_1$:$C_2$:...:$C_L$ where L is the maximum nesting level, automatically determined by the *Infomap* algorithm.

Table 1. 2-digit *Unesco* taxonomy terms

| | | | |
|---|---|---|---|
| 11 - Logic | 12 - Mathematics | 21 – Astronomy, Astrophysics | 22 – Physics |
| 23 - Chemistry | 24 - Life Sciences | 25 – Earth and Space Science | 31 – Agricultural Sciences |
| 32 – Medical Sciences | 33 – Technological Sciences | 51 - Anthropology | 52 - Demography |
| 52 – Economic Sciences | 54 - Geography | 55 - History | 56 – Juridical Science and Law |
| 57 - Linguistics | 58 - Pedagogy | 59 – Political Science | 61 – Psychology |
| 62 – Sciences of Arts and Letters | 62 - Sociology | 71 - Ethics | 72 – Philosophy |

### 2.3. Mapping to existing categories

Once the internal structure of the graph is revealed, it can be mapped to the existing categories used to describe pages in each detected community. Nevertheless, it is not possible to use *Wikipedia* categories as they are neither mutually exclusive nor collectively exhaustive (**Salah** *et al.*, 2012). Therefore, we decided to take advantage of a popular taxonomy used to categorize scientific (in a wide sense) content, namely the '*Unesco* nomenclature for fields of science and technology' (*Unesco*, 1988). This taxonomy provides three levels of refinement through a two, four, and six digit-based scheme for coding knowledge domains. There are 24 two-digit taxonomy terms (shown in Table 1), 245 four-digit taxonomy terms, and 2183 six-digit taxonomy terms in the 1988 version. Despite its obsolescence (*Unesco* abandoned it in 1992), representativeness and usability problems (**Martínez-Frías**; **Hochberg**, 2007), it has become a *de facto* standard for categorizing scientific and technological knowledge, and it constitutes a good starting point (**Ruiz-Martínez**; **Baños-Moreno**; **Martínez-Béjar**, 2014) for many other purposes.

It is not surprising that the *Unesco* taxonomy also has a page in *Wikipedia*, although the coverage is different across language versions. In the case of the Spanish *Wikipedia*, it is possible to retrieve all taxonomy terms from a single entry point, namely the six-digit taxonomy page. For instance, the first term is '11 (Logic)', which links to a *Wikipedia* page containing a link for such page and links to other six four-digit categories, namely '1101 (Application of Logic)', '1102 (Deductive Logic)', '1103 (General Logic)', '1104 (Inductive Logic)', '1105 (Methodology)', and '1199 (Other specialties relating to Logic)'. Four of these four-digit categories also link to *Wikipedia* pages, although one of them is empty. Within each four-digit term there are zero or more six-digit taxonomy terms, such as '110201 (Analogy)', for instance, which points to a *Wikipedia* page defining such concept. Unfortunately, this is different in languages other than Spanish, where these pages may not exist. This is a perfect example of the different levels of development that can be found in *Wikipedia* (**Halavais**; **Lackaff**, 2008; **Jiménez-Pelayo**, 2009; **Samoilenko** *et al.*, 2016), as content is created according to users' interests, with little coordination.

In the Spanish case, the top-level *Unesco* taxonomy page points to a total of 1,251 *Wikipedia* pages, although unfortunately some of them are empty web pages, point to wrong pages, or to repeated ones. Once cleansed, the *Wikipedia Unesco* taxonomy points to 974 valid *Wikipedia* pages, that is, 44.6% of the total six-digit *Unesco* categories.

## 3. Results

We have used the available *Infomap* implementation as described in **Bohlin** *et al.* (2014), using default values for the input parameters. *Infomap* is known to be one of the best algorithms for this purpose (**Lancichinetti**; **Fortunato**, 2009). Nevertheless, as stated in **Emmons** *et al.* (2016), *Infomap* can fail to identify large communities in very large networks, as in our case. On the other hand, *Infomap* may also fail to generate accurate communities if there are too many small ones, as also occurs in our case. These two issues mean that:

a) several pages from the same topic might fall in different communities; and

b) small communities might not be detected at all.

### 3.1. Communities detected

We obtained 35,296 communities and subcommunities (up to five nesting levels), containing between 1 and 17,588 pages, following a long-tail distribution, as only 1,442 communities included 100 or more *Wikipedia* pages, containing 79.1% of all the *Wikipedia* pages used to create the graph. If we want to include at least 90% of all the *Wikipedia* pages in the analysis, we would need to analyze 3,662 communities containing 32 or more pages. In theory, each one of these 3,662 communities should be manually inspected, in order to determine the main topic. For instance, the largest community (coded by the *Infomap* algorithm as '1:1') contains 17,588 pages about movies, actors and actresses, directors, and cinema in general. The second one (coded as '1:15') contains 13,633 pages about Chile, the country. On the other hand, the third largest community (coded as '2:2:6') contains 13,420 pages about the order 'Coleoptera', which can be considered scientific content. Obviously, manually examining 3,662 communities is not reasonable, so we need to use another approach that reduces the number of communities to inspect to a reasonable amount. **Figuerola**, **Groves** and **Quintanilla** (2015) describe a similar approach, but they only analyze the 255 largest top-level communities detected by *Infomap* in an older dump (November 2013), without describing other details of the manual review process. Furthermore, they use the *Wikipedia* dump without removing special pages such as categories, which were discarded in our analysis because of the aforementioned inconsistencies. Table 2 summarizes the structure of the most important communities and subcommunities for the top levels. The distribution of communities found by **Figuerola**, **Groves,** and **Quintanilla** (2015) was completely different, although the authors also identified long tails with very small communities.

Instead of analyzing each community and trying to determine whether it can be considered part of the science and technology corpus or not, following a top-down approach, we will use a bottom-up approach, starting from the *Unesco* six-digit taxonomy pages. The other pages in the same community will then also be assumed to belong to this category, bounding the probability of misclassification by means of sampling analysis.

Table 2. Top-level communities and subcommunities detected by *Infomap*

| Community | Pages | Subcommunities | Largest subcommunities | Topic |
|---|---|---|---|---|
| 1 | 1,013,758 (88.7%) | 12,054 | 1:1 (17,588) | Movies |
| | | | 1:15 (13,633) | Chile |
| | | | 1:12 (11,453) | Soccer |
| | | | 1:10 (10,144) | Russia |
| | | | 1:19 (9,866) | Music |
| 2 | 90,363 (7.9%) | 637 | 2:2:6 (13,420) | Coleoptera |
| | | | 2:2:7 (3,360) | Coleoptera |
| | | | 2:15 (2,949) | Gastropoda |
| | | | 2:35 (1,780) | Hemiptera |
| | | | 2:2:12 (1,731) | Coleoptera |
| 3 | 12,084 (1.1%) | 725 | 3:2 (393) | Bacteria |
| | | | 3:1 (299) | Cell |
| | | | 3:9 (197) | Psychiatry |
| | | | 3:14 (189) | Cancer |
| | | | 3:3 (167) | AIDS |
| 4 | 21,768 (1.9%) | 3,719 | 4:18 (550) | Siluriformes |
| | | | 4:11 (490) | Perciformes |
| | | | 4:9 (463) | Cypriniformes |
| | | | 4:1 (442) | Fishes (general) |
| | | | 4:5 (432) | Characiformes |
| 5-200 | 495 (< 0.4%) | NA | NA | NA |

## 3.2. Mapping to the *Unesco* taxonomy

The *Unesco* nomenclature is present in *Wikipedia* as a three-level hierarchical taxonomy, as described in **Ruiz-Martínez**, **Baños-Moreno** and **Martínez-Béjar** (2014). There is a *Wikipedia* page for two-digit *Unesco* terms pointing to other pages containing four-digit terms, which in turn point to pages containing six-digit terms. These *Wikipedia* pages also include links to other *Wikipedia* pages with content describing the concepts used in the *Unesco* nomenclature. Therefore, starting from the *Wikipedia* page describing the two-digit *Unesco* nomenclature, 974 different *Wikipedia* pages can be reached.

For each one of these 974 *Wikipedia* pages, we have identified the community that the page belongs to, using the results provided by the *Infomap* algorithm, which assigns a community for each page. These 974 pages can be mapped to 458 different communities, containing a total of 205,907 *Wikipedia* pages, that is, 18.0% of the total number of *Wikipedia* pages considered to be content articles. Nevertheless, we must ensure that all pages reachable from the *Unesco* six-digit taxonomy page are truly scientific and technological content. For instance, all communities containing pages related to artistic works (music, paintings, etc.) or geographical places were discarded.

> The *Unesco* nomenclature is present in *Wikipedia* as a three-level hierarchical taxonomy

For this purpose, the 458 identified communities were then manually inspected in order to determine whether they could really be considered scientific and technological content or not. We followed a procedure derived from a single-sampling acceptance plan (**Ruggeri**; **Kenett**; **Faltin**, 2007), treating communities as 'lots', but taking into account the large variability of community size, thus establishing criteria for small and large communities. In order to do so, each community was inspected by two researchers (out of five), according to the following criteria:

- If the community was small (i.e. less than or equal to 125 articles, a total of 302 communities out of 458), it was completely screened by searching for content not considered to be scientific or technological content, measuring the percentage of false positives. Communities with more than 15% of false positives were rejected.
- If the community was too large for complete screening, a tag cloud containing all the words in the pages' titles was built for preliminary analysis purposes, detecting communities with unexpected terms that could be rapidly discarded. For instance, communities containing pages about cities or geographical regions fell into this category. Additionally, a subset of 125 articles was further inspected, discarding the community if more than 20% of false positives was found. This is equivalent to an inspection level II single sampling acceptance plan.
- If the two researchers did not agree on whether a community should belong to the scientific and technological corpus or not, it was then inspected once again by all the researchers, reaching a consensus decision.
- Decisions were made at the community level, so no individual pages were discarded or included in the corpus.

This process generated a final corpus of 340 communities containing a total of 60,108 *Wikipedia* pages. This is only 5.3% of the total number of *Wikipedia* pages considered, but all of them can be considered to be content related to science and technology, with almost no false positives. After cleansing, **Figuerola**, **Groves** and **Quintanilla** (2015) found 119,797 *Wikipedia* pages that were considered to contain scientific and technological content (11.7%), although the

Table 3. Number of pages belonging to each *Unesco* two-digit taxonomy term

| ***Unesco* 2-digit term** | 33 | 24 | 21 | 25 | 22 | 12 | 23 | 63 |
|---|---|---|---|---|---|---|---|---|
| **Number of pages** | 13,986 | 8,598 | 8,253 | 5,386 | 4,329 | 3,829 | 2,425 | 2,260 |
| **%** | 23.3 | 14.3 | 13.7 | 9.0 | 7.2 | 6.4 | 4.0 | 3.8 |
| ***Unesco* 2-digit term** | 72 | 32 | 53 | 61 | 31 | 57 | 51 | 59 |
| **Number of pages** | 1,936 | 1,701 | 1,604 | 985 | 985 | 928 | 900 | 744 |
| **%** | 3.2 | 2.8 | 2.7 | 1.6 | 1.6 | 1.5 | 1.5 | 1.2 |
| ***Unesco* 2-digit term** | 55 | 11 | 52 | 56 | 54 | 71 | 58 | 62 |
| **Number of pages** | 344 | 336 | 297 | 144 | 77 | 61 | 0 | 0 |
| **%** | 0.6 | 0.6 | 0.5 | 0.2 | 0.1 | 0.1 | 0.0 | 0.0 |

authors did not bound the possible misclassification error when considering each community as a whole.

In light of these results and answering question Q1, we can say that it is possible to extract scientific and technological content without using the existing category structure, and we provide a method for this task.

It is remarkable that from the 24 *Unesco* two-digit taxonomy terms (shown in Table 1), 22 appear in the corpus. The two terms that do not appear in the corpus are '58 (Pedagogy)' and '62 (Sciences of Arts and Letters)'. The latter was not considered to be part of the corpus, as previously mentioned. On the other hand, the former is a *Wikipedia* page with no links to other pages, thus becoming an isolated page not traversed by the *Infomap* algorithm and, therefore, not belonging to any relevant community. On the other hand, as shown in Table 3, the six most represented *Unesco* two-digit taxonomy terms are 'Technological Sciences' (33), 'Life Sciences' (24)', 'Astronomy, Astrophysics' (21), 'Earth and Space Sciences' (25), 'Physics' (22) and 'Mathematics' (12), accounting for more than 73.9% of the total pages considered scientific and technological content.

Figure 1 shows a graph describing how *Unesco* communities are related to each other according to the number of links they share, that is, their similarity in a distance sense. This graph was created with *VOS* (**Van-Eck**; **Waltman**, 2007). Nodes represent *Unesco* categories and edges represent links between pages belonging to two categories. Node size is proportional to the number of pages in that category, while edge width is proportional to the number of links shared between those communities. *VOS* identifies three large clusters of communities, each one of them coded in a different color.

Notice that there is a clear bias towards technological content (the largest nodes), while social sciences or medicine are clearly underrepresented, as stated by **Halavais** and **Lackaff** (2008) and more recently in **Mesgari** *et al.* (2015). Ne-

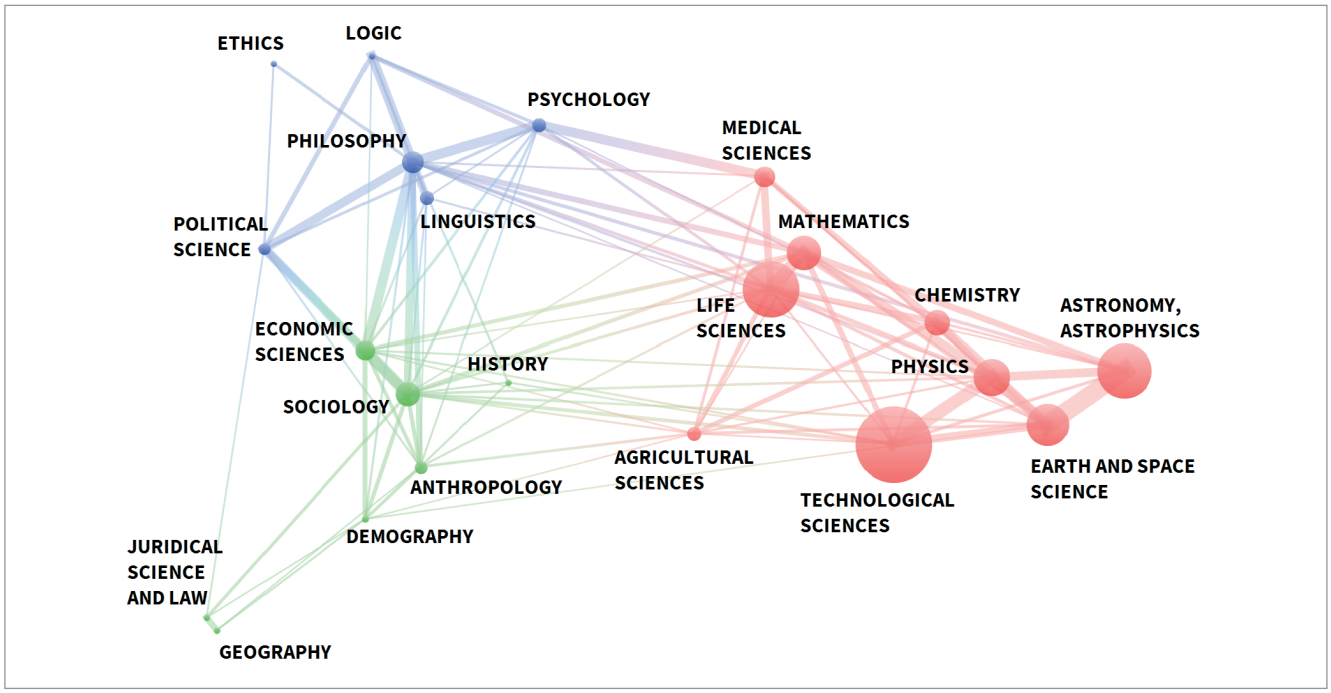



Figure 1. *VOS* map of *Unesco* communities according to the number of links they share within the scientific and technological corpus.

vertheless, there are some similarities with standard cognitive maps of science, as the one proposed by **Rafols**, **Porter** and **Leydesdorff** (2010): a quasi-modular and fragmented structure with some noticeable clusters and without a clear center. The clusters identified by *VOS* are also comparable to those described in **Rafols**, **Porter** and **Leydesdorff** (2010), as well as the main edges between the most important nodes within and between clusters. For instance, there is a strong edge connecting "Astronomy, Astrophysics" with "Earth and Space Science" and another one connecting "Psychology" with "Medical Sciences".

Therefore, answering question Q2, our method produces a comprehensive corpus across disciplines and quite consistent with contemporary global images of science, although it also reflects the known bias of *Wikipedia* towards technological content.

> *Wikipedia* has a clear bias towards technological content (the largest nodes), while social sciences or medicine are clearly underrepresented

## 4. Conclusions

*Wikipedia* has recently become –among many other things- the main source of scientific information for the general public and it is increasingly used in academia, by students and faculty (**Aibar** *et al.*, 2015). Recent studies on communication and public perception of science in several countries agree that the Internet has turned out to be, for most people, the main source of scientific information and that within the Internet, *Wikipedia* is the single most browsed site. People use it not only to satisfy their curiosity on the Higgs boson or on special relativity, but also to learn about more 'sensitive' issues of health and medical conditions that might be used to make practical decisions. Nonetheless, scientists and scientific institutions do not seem to be aware of this situation. Therefore, it is important to analyze the nature of scientific and technological content in *Wikipedia*, which needs to be identified by other means than the current categories, which have not been systematically created.

Taking advantage of the highly interlinked structure of *Wikipedia*, we used a graph random walk algorithm for detecting communities of pages strongly linked to each other. The original *Wikipedia* dump was pre-processed to generate a graph, removing duplicates, special pages, and resolving language related issues. A total of 60,108 *Wikipedia* pages distributed in 340 communities were identified as scientific and technological content. These pages belong to those communities that are reachable from the 974 *Unesco* six-digit taxonomy terms, removing those communities that cannot be considered scientific or technological. From the original 24 categories in the *Unesco* two-digit taxonomy, 22 are represented in the corpus, with a clear bias towards technology and natural sciences. The resulting graphical layout of the relationships between the different fields –as a function of the intern links between the articles within them- shows a coherent picture when compared to standard cognitive maps of science and a high correlation with other databases such as *Scopus* and *Web of Science*.

This small subset (5.3% of the total Spanish *Wikipedia* pages) is, nevertheless, a valuable resource for analyzing differences in *Wikipedia* as a mechanism for transmitting scientific and technological content compared to common knowledge. It also shows the large amount of scientific and technological content in *Wikipedia* that can be semi-automatically discovered, which is probably larger and could be extended by other means. *Wikipedia* is, therefore, a great portal for introducing people to scientific and technological content, establishing a bridge between society and academia.

Besides obtaining the scientific and technological corpus, the results of this analysis could be directly applied to *Wikipedia*, as up to three new categories could be added to all the pages considered to be part of it, describing the *Unesco* two-digit, four-digit, and six-digit taxonomies that the page belongs to. On the other hand, some of the detected communities have no link to the *Unesco* taxonomy pages, so they could also be improved by adding these new terms. Finally, the *Unesco* taxonomy could be recreated as a *Wikipedia* category structure and linked to the general "Science" category, helping *Wikipedia* users to find scientific and technological content. Providing a better support for categories would encourage increased participation in the expansion and refinement of the category system, especially among novice editors, as stated in **Thornton** & **McDonald** (2012).

> The *Unesco* taxonomy could be recreated as a *Wikipedia* category structure and linked to the general "Science" category, helping *Wikipedia* users to find scientific and technological content

Current and future research on this topic should take into account two important dimensions in *Wikipedia*: temporal, as it is an evolving environment, thus promoting longitudinal studies; and idiomatic, as different languages may show distinct cultural approaches towards scientific and educational content. Once a scientific and technological corpus has been established, it is possible to analyze, among other things, whether *Wikipedia* shows an accurate representation of present scientific knowledge –comparing the distribution of articles with standard repositories of scientific output- or not, as well as whether mainstream positions tend to be favored in dealing with controversial issues, and the kind of sources that are used by editors.

## Acknowledgements

This research is part of the project "Análisis del contenido científico de la *Wikipedia* española", funded by *Fecyt*, ref. FCT-14-8269.

## 5. References

**Aibar, Eduard**; **Lladós-Masllorens, Josep**; **Meseguer-Artola, Antoni**; **Minguillón, Julià**; **Lerga, Maura** (2015). "Wikipedia

at university: what faculty think and do about it". *The electronic library*, v. 33, n. 4, pp. 668-683.
*http://openaccess.uoc.edu/webapps/o2/handle/10609/39442*
*https://doi.org/10.1108/EL-12-2013-0217*

**Barabási, Albert-László**; **Albert, Réka** (1999). "Emergence of scaling in random networks". *Science*, v. 286, pp. 509-512.
*http://barabasi.com/f/67.pdf*
*https://doi.org/10.1126/science.286.5439.509*

**Bohlin, Ludvig**; **Edler, Daniel**; **Lancichinetti, Andrea**; **Rosvall, Martin** (2014). "Community detection and visualization of networks with the map equation framework". In: Ding, Ying; Rousseau, Ronald; Wolfram, Dietmar (eds.). *Measuring scholarly impact*. Springer International Publishing, pp. 3-34.
*https://doi.org/10.1007/978-3-319-10377-8_1*

**Brossard, Dominique**; **Scheufele, Dietram** (2013). "Science, new media, and the public". *Science*, v. 339, n. 6115, pp. 40-41.
*https://goo.gl/9W5zZR*
*https://doi.org/10.1126/science.1232329*

**Capocci, Andrea**; **Rao, Francesco**; **Caldarelli, Guido** (2007). "Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia *Wikipedia*". *Europhysics letters*, v. 81, n. 2.
*https://arxiv.org/abs/0710.3058*
*https://doi.org/10.1209/0295-5075/81/28006*

**Emmons, Scott**; **Kobourov, Stephen**; **Gallant, Mike**; **Börner, Katy** (2016). "Analysis of network clustering algorithms and cluster quality metrics at scale". *PLoS one*, v. 11, n. 7, art. no. e0159161.
*https://doi.org/10.1371/journal.pone.0159161*

**Ermann, Leonardo**; **Frahm, Klaus M.**; **Shepelyansky, Dima L.** (2015). "*Google* matrix analysis of directed networks". *Reviews of modern physics*, v. 87, n. 4.
*https://arxiv.org/abs/1409.0428*
*https://doi.org/10.1103/RevModPhys.87.1261*

**Figuerola, Carlos G.**; **Groves, Tamar**; **Quintanilla, Miguel-Ángel** (2015). "The implications of *Wikipedia* for contemporary science education: Using social network analysis techniques for automatic organisation of knowledge". In: *Proceedings of the 3rd Intl conf on technological ecosystems for enhancing multiculturality*, TEEM'15, Porto, October 7-9, pp. 403-410.
*http://eprints.rclis.org/29277/1/20156_figue.pdf*
*https://doi.org/10.1145/2808580.2808641*

**Halavais, Alexander; Lackaff, Derek** (2008). "An analysis of topical coverage of *Wikipedia*". *Journal of computer-mediated communication*, v. 13, n. 2, pp. 429-440.
*https://doi.org/10.1111/j.1083-6101.2008.00403.x*

**Hejazy, Khaled A.**; **El-Beltagy, Samhaa R.** (2013). "An approach for deriving semantically related category hierarchies from *Wikipedia* category graphs". In: Rocha, Álvaro; Correia, Ana-María; Wilson, Tom; Stroetmann, Karl A. (eds.). *Advances in information systems and technologies*. Berlin Heidelberg: Springer, pp. 77-86.

**Holloway, Tod**; **Božicevic, Miran**; **Börner, Katy** (2007). "Analyzing and visualizing the semantic coverage of *Wikipedia* and its authors". *Complexity*, v. 12, n. 3, pp. 30-40.
*http://nwb.cns.iu.edu/papers/holloway-0000-analvizwikipedia.pdf*
*https://doi.org/10.1002/cplx.20164*

**Jiang, Yuncheng**; **Zhang, Xiaopei**; **Tang, Yong**; **Nie, Ruihua** (2015). "Feature-based approaches to semantic similarity assessment of concepts using *Wikipedia*". *Information processing & management*, v. 51, n. 3, pp. 215-234.
*https://doi.org/10.1016/j.ipm.2015.01.001*

**Jiménez-Pelayo, Jesús** (2009). "*Wikipedia* como vocabulario controlado: ¿está superado el control de autoridades tradicional". *El profesional de la información*, v. 18, n. 2, pp. 188-201.
*https://doi.org/10.3145/epi.2009.mar.09*

**Kaptein, Rianne**; **Kamps, Jaap** (2013). "Exploiting the category structure of *Wikipedia* for entity ranking". *Artificial intelligence*, v. 194, pp. 111-129.
*https://doi.org/10.1016/j.artint.2012.06.003*

**Kim, Kyung-Sun**; **Sin, Sei-Ching-Joanna**; **Yoo-Lee, Eun-Young** (2014). "Undergraduates' use of social media as information sources". *College & research libraries*, v. 75, n. 4, pp. 442-457.
*https://doi.org/10.5860/crl.75.4.442*

**Lancichinetti, Andrea**; **Fortunato, Santo** (2009). "Community detection algorithms: A comparative analysis". *Physical review E - statistical, nonlinear, and soft matter physics*, v. 80, n. 5, art. no. 056117.
*https://arxiv.org/abs/0908.1062*
*https://doi.org/10.1103/PhysRevE.80.056117*

**Lih, Andrew** (2004). "Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource". In: *Procs of the 5th Intl symposium on online journalism*, pp. 16-17.
*https://goo.gl/6XpLn9*

**Martínez-Frías, Jesús**; **Hochberg, David** (2007). "Classifying science and technology: two problems with the *Unesco* system". *Interdisciplinary science reviews*, v. 32, n. 4, pp. 315-319.
*http://digital.csic.es/handle/10261/13013*

**Medelyan, Olena**; **Milne, David**; **Legg, Catherine**; **Witten, Ian** (2009). "Mining meaning from *Wikipedia*". *International journal of human-computer studies*, v. 67, n. 9, pp. 716-754.
*https://arxiv.org/abs/0809.4530*
*https://doi.org/10.1016/j.ijhcs.2009.05.004*

**Mesgari, Mostafa**; **Okoli, Chitu**; **Mehdi, Mohamad**; **Nielsen, Finn-Årup**; **Lanamäki, Arto** (2015). "'The sum of all human knowledge': A systematic review of scholarly research on the content of *Wikipedia*". *Journal of the Association for Information Science and Technology*, v. 66, n. 2, pp. 219-245.
*http://spectrum.library.concordia.ca/978618/*
*https://doi.org/10.1002/asi.23172*

**Muchnik, Lev**; **Itzhack, Royi**; **Solomon, Sorin**; **Louzoun, Yoram** (2007). "Self-emergence of knowledge trees: Extraction of the *Wikipedia* hierarchies". *Physical review E*, v. 76, n. 1.
*https://doi.org/10.1103/PhysRevE.76.016106*

**Pons, Pascal**; **Latapy, Matthieu** (2005). "Computing communities in large networks using random walks". In: Yolum, Pinar; Güngör, Tunga; Gürgen, Fikret; Özturan, Can (eds.). *Procs of the 20th Intl symposium on computer and information sciences* (*ISCIS 2015*). Berlin, Heidelberg: Springer, pp. 284-293.

*https://arxiv.org/abs/physics/0512106*

**Ponzetto, Simone-Paolo**; **Strube, Michael** (2007). "Knowledge derived from Wikipedia for computing semantic relatedness". *Journal of artificial intelligence research*, v. 30, pp. 181-212.
*https://doi.org/10.1613/jair.2308*

**Rafols, Ismael**; **Porter, Alan L.**; **Leydesdorff, Loet** (2010). "Science overlay maps: A new tool for research policy and library management". *Journal of the American Society for information Science and Technology*, v. 61, pp. 1871-1887.
*http://www.leydesdorff.net/overlaytoolkit/overlaytoolkit.pdf*
*https://doi.org/10.1002/asi.21368*

**Rosvall, Martin**; **Bergstrom, Carl T.** (2008). "Maps of random walks on complex networks reveal community structure". In: *Procs of the National Academy of Sciences*, v. 105, n. 4, pp. 1118-1123.
*https://doi.org/10.1073/pnas.0706851105*

**Ruggeri, Fabrizio**; **Kenett, Ron S.**; **Faltin, Frederick** (eds.) (2007). *Encyclopedia of statistics in quality and reliability*. Wiley. ISBN: 978 0 470061572
*https://doi.org/10.1002/9780470061572*

**Ruiz-Martínez, Juana-María**; **Baños-Moreno, María-José**; **Martínez-Béjar, Rodrigo** (2014). "Nomenclatura Unesco: evolución, alcance y reutilización en clave ontológica para la descripción de perfiles científicos". *El profesional de la información*, v. 23, n. 4, pp. 383-392.
*https://doi.org/10.3145/epi.2014.jul.06*

**Ryu, Pum-Mo**; **Jang, Myung-Gil**; **Kim, Hyun-Ki** (2014). "Open domain question answering using Wikipedia-based knowledge model". *Information processing & management*, v. 50, n. 5, pp. 683-692.
*https://doi.org/10.1016/j.ipm.2014.04.007*

**Salah, Almila-Akdag**; **Gao, Cheng**; **Suchecki, Krzysztof**; **Scharnhorst, Andrea** (2012). "Need to categorize: A comparative look at the categories of universal decimal classification system and *Wikipedia*". *Leonardo*, v. 45, n. 1, pp. 84-85.
*https://arxiv.org/abs/1105.5912*
*https://doi.org/10.1162/LEON_a_00344*

**Samoilenko, Anna**; **Karimi, Fariba**; **Edler, Daniel**; **Kunegis, J**érôme; **Strohmaier, Markus** (2016). "Linguistic neighbourhoods: Explaining cultural borders on *Wikipedia* through multilingual co-editing activity". *EPJ data science*, v. 5, n. 9.
*https://doi.org/10.1140/epjds/s13688-016-0070-8*

**Silva, Filipi-Nascimento**; **Viana, Matheus-Palhares**; **Travençolo, Bruno A. N.**; **Costa, Luciano da F.** (2011). "Investigating relationships within and between category networks in *Wikipedia*". *Journal of informetrics*, v. 5, n. 3, pp. 431-438.
*https://goo.gl/6wF8cR*
*https://doi.org/10.1016/j.joi.2011.03.003*

**Snyder, Johnny** (2013). "*Wikipedia*: Librarians' perspectives on its use as a reference source". *Reference & user services quarterly*, v. 53, n. 2, pp. 155-163.
*https://doi.org/10.5860/rusq.53n2.155*

**Strogatz, Steven H.** (2001). "Exploring complex networks". *Nature*, n. 410, pp. 268-276.
*https://doi.org/10.1038/35065725*

**Suchecki, Krzysztof**; **Salah, Alkim-Almila-Akdag**; **Gao, Cheng**; **Scharnhorst, Andrea** (2012). "Evolution of Wikipedia's category structure". *Advances in complex systems*, v. 15 (supp01), 1250068.
*https://arxiv.org/abs/1203.0788*
*https://doi.org/10.1142/S0219525912500683*

**Thornton, Katherine**; **McDonald, David W.** (2012). "Tagging Wikipedia: collaboratively creating a category system". In: *Procs of the 17th ACM Intl conf on supporting group work*, pp. 219-228.
*http://www.pensivepuffin.com/dwmcphd/papers/Thornton_McDonald-TaggingWikipedia-GROUP12.pdf*
*https://doi.org/10.1145/2389176.2389210*

*Unesco* (1988). *Proposed international standard nomenclature for fields of science and technology*. NS/ROU/257 REV.1; SC.88/WS/80.
*http://unesdoc.unesco.org/images/0008/000829/082946eb.pdf*

**Van-Eck, Nees-Jan**; **Waltman, Ludo** (2007). "VOS: A new method for visualizing similarities between objects". In: Decker, Reinhold; Lenz, Hans J. (eds.). *Advances in data analysis: Procs of the 30th Annual conf of the German Classification Society*, Berlin, March 8-10. Springer, pp. 299-306.
*https://doi.org/10.1007/978-3-540-70981-7_34*

El profesional de la información

CRECS

ANUARIO think EPI

EPI scholar

DESAFÍO A LA INVESTIGACIÓN ESTÁNDAR EN COMUNICACIÓN
CRÍTICA Y ALTERNATIVAS
MANUEL GOYANES

El profesional de la información · EDITORIAL UOC

El profesional de la información

FUENTES DE INFORMACIÓN
GUÍA BÁSICA Y NUEVA CLASIFICACIÓN
ROSANA LÓPEZ-CARREÑO

EDITORIAL UOC

CroDoc

e-LiS

IRALIS®

IWETEL

COMUNICACION

INCYT

exit