Retrieval of very large numbers of items in the *Web of Science*: an exercise to develop accurate search strategies

By **Ricardo Arencibia-Jorge, Loet Leydesdorff, Zaida Chinchilla-Rodríguez, Ronald Rousseau** and **Soren W. Paris**

Ricardo Arencibia-Jorge, *máster en bibliotecología y ciencia de la información, es jefe del Departamento de Información Científica del Centro Nacional de Investigaciones Científicas (CNIC) de La Habana, Cuba. Desarrolla su investigación doctoral en el análisis cienciométrico de la actividad científica cubana. Coordinador del proyecto Red de Estudios Cienciométricos sobre la Educación Superior cubana (Redec). Editor para Cuba del repositorio de información E-LIS y premio de la Asociación Cubana de Bibliotecarios en 2009.*

Loet Leydesdorff, *doctor en sociología y profesor titular del Departamento de Dinámica de la Ciencia y la Tecnología en la Universidad de Amsterdam, Holanda, posee una vasta experiencia en diferentes campos como la sociología de la ciencia, el análisis de redes sociales y la sociología de la innovación. En el 2003 recibió el premio Derek de Solla Price por su contribución a la cienciometría. Autor de múltiples libros y artículos que abarcan los dominios de la ciencia de la información y la comunicación.*

Zaida Chinchilla-Rodríguez *es doctora en documentación e información científica y científica titular del Consejo Superior de Investigaciones Científicas (CSIC), en el Instituto de Políticas y Bienes Públicos (IPP) de Madrid, España. Miembro del Grupo de Investigación SCImago, desarrolla su investigación en el análisis cienciométrico de dominios del conocimiento, la representación y visualización de información y redes de colaboración científica, y propuestas metodológicas para el diseño de sistemas de información.*

Ronald Rousseau, *doctor en bibliotecología y ciencia de la información, profesor de la Escuela Católica de Educación Superior de Brujas-Ostende (KHBO), e investigador asociado de la Univ. Católica de Leuven, Bélgica. Presidente de la Sociedad Internacional de Cienciometría e Informetría (ISSI). Investiga sobre estudios métricos de la información y evaluación de la investigación. Ha recibido importantes galardones como el premio de la Academia de Ciencias de Bélgica en 1979, y el Derek J. de Solla Price en 2001.*

Abstract: *The Web of Science interface counts at most 100,000 retrieved items from a single query. If the query results in a dataset containing more than 100,000 items the number of retrieved items is indicated as >100,000. The problem studied here is how to find the exact number of items in a query that leads to more than 100,000 items. One way to achieve this objective is presented. The retrieval of the entire scientific production from the United States in a specific year (2007) is counted and an advanced search strategy is designed. Different sections of items can be retrieved using the Source field of the database. A Boolean statement was created with the aim of eliminating overlapped sections and improving the accuracy of this search strategy.*

Keywords: *Information retrieval, Search strategies, Databases, Web of Science, Scientific production, USA.*

Soren W. Paris, *licenciado en lengua inglesa por la Universidad de West Chester, Estados Unidos, en 2001. Actualmente cursa la Maestría en Gestión y Conservación de Recursos de la Universidad de Antioch, en New Hampshire. Desde el 2002, ha sido investigador asistente del Dr. Eugene Garfield en el área de tecnologías de información. Ambos han trabajado en el desarrollo del programa HistCite, una herramienta para el análisis bibliométrico y la visualización de los resultados de las búsquedas de información en el Web of Science.*

Título: **Recuperación de grandes cantidades de registros en la *Web of Science*: un ejercicio para realizar estrategias de búsqueda precisas**

Resumen: *La interfaz de la Web of Science permite recuperar como máximo 100.000 registros en una búsqueda simple. Si el resultado de la búsqueda tiene más de 100.000 registros, el número de registros recuperados se indica como >100.000. En este artículo se presenta una forma de encontrar el número total de registros en una búsqueda que supera los 100.000 registros. Concretamente, se contabiliza la producción científica total de los Estados Unidos en un año específico (2007). Se diseña una estrategia de búsqueda avanzada para recuperar conjuntos diferentes de registros usando el campo Source*

de la base de datos. Se crea una instrucción booleana con el fin de eliminar los solapos y mejorar la precisión de la estrategia de búsqueda.

Palabras clave: *Recuperación de información, Estrategias de búsqueda, Bases de datos,* Web of Science*, Producción científica, USA.*

Arencibia-Jorge, Ricardo; **Leydesdorff, Loet**; **Chinchilla-Rodríguez, Zaida**; **Rousseau, Ronald**; **Paris, Soren W**. "Retrieval of very large numbers of items in the Web of Science: an exercise to develop accurate search strategies". *El profesional de la información*, 2009, septiembre-octubre, v. 18, n. 5, pp. 529-533.

DOI: 10.3145/epi.2009.sep.06

1. Introduction

SOMETIMES INFORMATION PROFESSIONALS FACE SINGULAR PROBLEMS related to the use of information technology and the management of digital environments. Changes and improvements offered by online providers present users with new tools and different interfaces, requiring continual re-learning (Martínez, 2008).

Often an apparently simple and easy activity requires the practical knowledge of specialists. As many retrieval tasks are team work, each member of the team must clearly communicate objectives, solutions and experiences to the rest. Such working habits lead to a global increase in knowledge and skills.

"*WoS* data related to geopolitical domains with large numbers of items must be searched using a combination of search statements"

The problem presented here came up during a work session of the *SCImago* research team. Specialists from the Spanish group were doing a scientometric study of the world scientific production in 2007 using *Scopus* and *Web of Science* (*WoS*) interfaces, when they noticed an inconvenience which at first sight appeared easy to resolve. Researchers needed an accurate number of papers produced in the USA and the United Kingdom, but a precise number over 100,000 items using the *WoS* interface was not available.

Recently, one of the multiple papers of **Péter Jacsó** on search strategies and techniques in the most widely used citation-enhanced databases called attention to this topic (**Jacsó**, 2009). *WoS* data related to geopolitical domains with large numbers of items must be searched using a combination of search statements. The clearest examples were countries such as the USA or the United Kingdom, or blocks of countries such as the European Union, with a scientific production in mainstream journals of over 100,000 articles during a year.

The identification of items from the United Kingdom does not present major difficulties. The construction of two statements including and excluding the word "London" in the *Author Address* field can easily solve the problem. For example, using all databases comprising the *WoS* (*SCI-Expanded*, *SSCI*, *A&HCI*, *CPCI-S*, *CPCI-SSH*, *IC*, *CCR-Expanded*), and selecting all years in *Timespan*, a user can obtain the total output of this nation through the sum of the items retrieved by the following search statements:

1. PY=2007ANDCU=(England OR Scotland OR Wales OR North Ireland) AND AD=LONDON

2. PY=2007ANDCU=(England OR Scotland OR Wales OR North Ireland) NOT AD=LONDON

"To restrict the search to the year an article is published, it is necessary to use the *Publication Year* (PY) field"

Note that using the *Timespan Limits* the user is in fact restricting the search to the year the data were entered into the database. Therefore, to restrict the search to the year an article is published, it is necessary to use the *Publication Year* (PY) field.

As of June 18, 2009 (the date of this query), there were 33,043 articles in 2007 signed by authors from at least one London scientific or scholarly institution, and there were 98,802 in which there was no author from this English city. A total of 131,845 articles compose the sum total output of the United Kingdom in the *WoS* that year.

But, what about the USA? The scientific production from this country in a year far exceeds 100,000 articles. How to obtain the total output of the USA using the *WoS* interface? That question gave rise to a practical and interesting exercise, which required the united efforts of various specialists from different research groups.

2. In search of a solution

At first, a series of search strategies developed by the *SCImago* group was oriented towards the identification of the states of the Union in the *Author Address* (AD) field, with the aim of obtaining different sections of fewer than 100,000 items. But the design of this kind of advanced search strategies, based on the AD field, became very complex in this case. The extensive collaboration between institutions from different states made it difficult to construct a logical operation in the search strategy that would eliminate duplicates. **Ronald Rousseau** devised the most complete strategy, but the results required a very complex validation process. All the strategies and results were sent to **Eugene Garfield** and his assistant **Soren W. Paris**, who validated the results with their own results obtained from their direct searches in *Thomson Reuters* databases. In this case, there were still significant differences between the AD-based search strategy and the statistics compiled by **Paris**.

Based on previous personal experiences, **Loet Leydesdorff** proposed the use of a less problematical field to develop the search strategy: the *Source* (SO) field (**Zhou & Leydesdorff**, 2006). Thus, using the initial of the journal/proceedings title plus an asterisk (a truncation designed to retrieve all titles with the selected initial), the process of division into sections of fewer than 100,000 items was effective. The only problem was the existence of journals belonging to series, which were retrieved not only by the journal title, but also by the series title. In any case, there were only two possibilities to obtain duplicated data; that is, a journal could be covered by no more than two sections of fewer than 100,000 items. For this purpose, a Boolean statement in the search strategy with the aim to eliminate duplicates could be developed. **Leydesdorff**'s proposal was further developed by the *SCImago* research group, which finally devised a more accurate search strategy and developed the validation procedure.

3. Proposed search strategy

Table 1 shows the complete procedure devised to obtain the to-

Search strategy	Items	Sum
1. PY=2007 AND CU=USA AND (SO=A* OR SO=B*)	91,122	91,122
2. PY=2007 AND CU=USA AND (SO=C* OR SO=D* OR SO=E* OR SO=F* OR SO=G*)	91,920	183,042
3. PY=2007 AND CU=USA AND (SO=H* OR SO=I* OR SO=K* OR SO=L* OR SO=M*)	82,897	265,939
4. PY=2007 AND CU=USA AND (SO=N* OR SO=O* OR SO=P* OR SO=Q* OR SO=R*)	84,783	350,722
5. PY=2007 AND CU=USA AND (SO=S* OR SO=T* OR SO=U* OR SO=V* OR SO=W* OR SO=X* OR SO=Y* OR SO=Z* OR SO=1* OR SO=2* OR SO=3* OR SO=4* OR SO=5* OR SO=6* OR SO=7* OR SO=8* OR SO=9*)	58,751	409,473
6. PY=2007 AND CU=USA AND SO=J* AND AD=CA	17,064	426,537
7. PY=2007 AND CU=USA AND SO=J* NOT AD=CA	92,976	**519,513**
Statement to find overlapping	**Items**	**Sum**
8. (#1 AND #2) OR (#1 AND #3) OR (#1 AND #4) OR (#1 AND #5) OR (#1 AND #6) OR (#1 AND #7) OR (#2 AND #3) OR (#2 AND #4) OR (#2 AND #5) OR (#2 AND #6) OR (#2 AND #7) OR (#3 AND #4) OR (#3 AND #5) OR (#3 AND #6) OR (#3 AND #7) OR (#4 AND #5) OR (#4 AND #6) OR (#4 AND #7) OR (#5 AND #6) OR (#5 AND #7) OR (#6 AND #7)	**23,026** (Overlapping)	**496,487** (Σ 1-7) - 8
New Search Strategy (Excluding overlapping)	**Items**	**Sum**
9. #1 NOT #8	85,586	85,586
10. #2 NOT #8	87,535	173,121
11. #3 NOT #8	69,457	242,578
12. #4 NOT #8	75,516	318,094
13. #5 NOT #8	45,551	363,645
14. #6 NOT #8	17,008	380,653
15. #7 NOT #8	92,808	473,461
Sum 9-15 plus articles excluded by overlapping		**496,487** (Σ 9-15) + 8

Table 1. Search strategy to obtain the total number of articles from the United States of America in 2007 through the WoS interface (Databases = SCI-Expanded, SSCI, A&HCI, CPCI-S, CPCI-SSH; Timespan = All years; All document types). May 18, 2009

tal number of articles produced by institutions from the United States.

The first 7 statements were created with the aim of dividing the results into sections of fewer than 100,000 items. In each statement, the necessary journal initials, in alphabetical order, to obtain an upper limit of fewer than 100,000 results were used. Note that statements #6 and #7 were shaped with the same philosophy as the United Kingdom output retrieval procedure. There were more than 100,000 USA articles published in journals whose titles begin with "*J*". Therefore, the AD field was used to divide this specific section in two: articles published in these journals including authors belonging to institutions from California (CA), and excluding them. In the end, a total number of 519,513 articles was obtained.

Then, a Boolean statement (#8) was created to identify overlapping and to improve the accuracy of the search strategy. Removing these 23,026 overlapping items from the previously calculated number, a final number of 489,487 items was obtained.

With the purpose of identifying inaccuracies in this calculation process, the first 7 statements were implemented again (#9 to #15), but excluding items in overlapping sections. This gave a result of 473,461 items. The items in the overlapping sections were added, and 489,487 items were once again obtained. This number established a hypothetical total number of articles published by institutions from the USA during the year 2007.

4. Validation process

The total scientific output of a less productive country than the USA or the United Kingdom was tested. This could have been any nation from the rest of the world, but we used Cuba as a test case. A direct method was used to find the Cuban scientific production in *WoS* during the year 2007 (Table 2).

Search strategy	Items	Sum
1. PY=2007 AND CU=CUBA	910	**910**

Table 2. Search strategy to obtain the total number of Cuban articles in 2007 through the WoS interface: direct method (Databases = SCI-Expanded, SSCI, A&HCI, CPCI-S, CPCI-SSH; Timespan = All years; All document types)

A total of 910 items were identified using the word "Cuba" in the *Affiliation Country* (CU) field. So, the second step was to use the same strategy as the one developed to retrieve the total USA output. If the search strategy was correctly developed, the final number obtained by either of the two indirect methods (including and excluding overlapping sections) had also to be precisely 910. The table 3 confirms, finally, the accuracy of data obtained through the search strategy developed during the exercise.

Furthermore, results obtained from the *WoS* following this search strategy were in complete accordance with results reported independently by the *Thomson Reuters* team.

5. Final considerations

This exercise provided a methodology to obtain the same result in two different ways: a) searching with overlapping, and subtracting items in overlapping sections at the end; and b) searching without overlapping and adding the items in overlapping sections at the end. The use of a small country during the validation procedure allowed us to obtain the same total number not only through the proposed strategies, but also using a direct method, confirming the accuracy of the results and the efficacy of the search method.

This kind of SO-based search strategy is probably not the only alternative to retrieve USA scientific production in the *WoS*, and it may be that its implementation does not solve other problems related to large numbers of items to be retrieved using the *WoS* interface. In any case, for scientometric purposes, a fast and well described method to obtain reliable data is always welcome. In this sense, the method devised is an accurate and validated search strategy to be used by any specialist around the world, and the procedure presented shows the importance of team work in the development of advanced search strategies for information retrieval.

6. Acknowledgments

To **Eugene Garfield**, for all the support and advices. To **Félix de Moya Anegón**, **Carmen López Illescas**, **Elena Corera Álvarez**, **María Benavent Pérez** (*SCImago Research Group*, *Institute of Public Goods and Policies*, *CSIC*), for the team work. *Thomson Reuter*'s databases were available in Spain thanks to the *Spanish Foundation for Science and Technology* and the *Ministry of Science and Innovation* of the Spanish government.

7. References

Jacsó, Péter. "Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries". *Online information review*, 2009, v. 33, pp. 376-385.

Martínez, Luis-Javier. "La nueva versión de ISI Web of Knowledge: calidad y complejidad". *El profesional de la información*, 2008, v. 17, pp. 331-339.

Zhou, Ping; **Leydesdorff, Loet**. "The emergence of China as a leading nation in science". *Research policy*, 2006, v. 35, pp. 83-104.

Ricardo Arencibia-Jorge[a,c], ***Loet Leydesdorff***[b], ***Zaida Chinchilla-Rodríguez***[c], ***Ronald Rousseau***[d,e], *and* ***Soren W. Paris***[f]

Search strategy	Items	Sum
1. PY=2007 AND CU=CUBA AND (SO=A* OR SO=B*)	140	140
2. PY=2007 AND CU=CUBA AND (SO=C* OR SO=D* OR SO=E* OR SO=F* OR SO=G*)	216	356
3. PY=2007 AND CU=CUBA AND (SO=H* OR SO=I* OR SO=K* OR SO=L* OR SO=M*)	161	517
4. PY=2007 AND CU=CUBA AND (SO=N* OR SO=O* OR SO=P* OR SO=Q* OR SO=R*)	193	710
5. PY=2007 AND CU=CUBA AND (SO=S* OR SO=T* OR SO=U* OR SO=V* OR SO=W* OR SO=X* OR SO=Y* OR SO=Z* OR SO=1* OR SO=2* OR SO=3* OR SO=4* OR SO=5* OR SO=6* OR SO=7* OR SO=8* OR SO=9*)	91	801
6. PY=2007 AND CU=CUBA AND SO=J* AND AD=Havana	108	909
7. PY=2007 AND CU=CUBA AND SO=J* NOT AD=Havana	35	944
Statement to find overlapping	**Items**	**Sum**
8. (#1 AND #2) OR (#1 AND #3) OR (#1 AND #4) OR (#1 AND #5) OR (#1 AND #6) OR (#1 AND #7) OR (#2 AND #3) OR (#2 AND #4) OR (#2 AND #5) OR (#2 AND #6) OR (#2 AND #7) OR (#3 AND #4) OR (#3 AND #5) OR (#3 AND #6) OR (#3 AND #7) OR (#4 AND #5) OR (#4 AND #6) OR (#4 AND #7) OR (#5 AND #6) OR (#5 AND #7) OR (#6 AND #7)	34 Overlapping	**910** (Σ 1-7) - 8
New search strategy (excluding overlapping)	**Items**	**Sum**
9. #1 NOT #8	127	127
10. #2 NOT #8	205	332
11. #3 NOT #8	139	471
12. #4 NOT #8	177	648
13. #5 NOT #8	86	734
14. #6 NOT #8	108	842
15. #7 NOT #8	34	876
Sum 9-15 plus articles excluded by overlapping		**910** (Σ 9-15) + 8

Table 3. Search strategy to obtain the total number of Cuban articles in 2007 through the WoS interface: indirect methods (Databases = SCI-Expanded, SSCI, A&HCI, CPCI-S, CPCI-SSH; Timespan = All years; All document types). May 19, 2009.

[a] National Scientific Research Center CNIC, Avenue 25 and 158 street, AP 6414 Havana, Cuba.

[b] Amsterdam School of Communication Research, University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands.

[c] CSIC, Institute of Public Goods and Policies, SCImago Research Group, Albasanz 26-28, 28037 Madrid, Spain.

[d] K.U.Leuven, Department of Mathematics, Celestijnenlaan 200B, 3001 Leuven (Heverlee), Belgium.

[e] *KHBO, Department Industrial Sciences and Technology, Zeedijk 101, 8400 Oostende, Belgium.*

[f] *The Scientist, 400 Market St. Philadelphia, PA 19106 USA.*

Address for correspondence:

***Ricardo Arencibia-Jorge**, National Scientific Research Center, Avenue 25 and 158 street, AP 6414 Havana, Cuba.*
ricardo.arencibia@cnic.edu.cu