Tratamiento de los nombres españoles en las bases de datos internacionales: su incidencia en la recuperación de información y en los análisis bibliométricos

Rafael Ruiz Pérez, Emilio Delgado López-Cózar
Elena Corera, Mª José Álvarez Suárez, Evaristo Jiménez Contreras

Universidad de Granada. Facultad de Biblioteconomía y Documentación
Grupo de Trabajo "Evaluación y Transferencia de la Producción Científica"

Resumen

Justificación. La escasa normalización que presentan los autores españoles en las bases de datos bibliográficas nacionales e internacionales —no catalográficas—, es un problema apuntando pero no investigado en las metodologías de los estudios bibliométricos y en los trabajos sobre recuperación de información.

Objetivos. En este trabajo presentamos los primeros resultados de una investigación más amplia que tiene como objetivos: 1) cuantificar la inconsistencia en el tratamiento de los autores españoles, 2) conocer los mecanismos que originan esta inconsistencia en los procesos de indización, 3) proponer recomendaciones y pautas de comportamiento correctoras para mejorar la normalización de los nombres personales españoles en las bases de datos internacionales.

Metodología. Se estudia una muestra de autores españoles modernos presentes en tres bases de datos, representativa de las estructuras más comunes de nombre español.

Resultados. Los resultados muestran que más del 50% de los autores españoles del campo médico presentan dos o más variantes en su forma del nombre; que el número de variantes de un autor se incrementa considerablemente en relación con su productividad en la base de datos y que los comportamientos de las tres bases de datos son similares.

Conclusiones. La magnitud del problema descrito tiene su origen en: 1) las variaciones del nombre que un autor utiliza en su producción científica para firmar sus trabajos, reflejadas en las fuentes (revistas) de las que emanan las referencias bibliográficas, 2) falta de rigor y homogeneidad en los procesos de tratamiento de la información por las BD. Por todo ello, es necesario iniciar un proceso de medidas correctoras.

1. INTRODUCCIÓN

Las bases de datos bibliográficas se han convertido en intermediarios imprescindibles en el proceso de transferencia de la información científica. La búsqueda, identificación, recuperación, localización y obtención de documentos pasa necesariamente por ellas. También, gracias a las facilidades que ofrecen para el tratamiento y recuperación masiva de información, son fuentes de referencia imprescindibles para la realización de estudios bibliométricos (Hawkins 1977).

Por ello, la calidad de los registros bibliográficos de dichas BD es un elemento capital para que, de un lado, la comunicación científica funcione de manera eficaz, de otro, para que los análisis bibliométricos posean un mínimo grado de fiabilidad y validez. En el primer caso, los errores y la inconsistencia en los registros tienen funestas consecuencias en la ejecución de las búsquedas —pérdida de documentos relevantes— y en el acceso a los documentos (Bell & Speer 1988). En el segundo, la calidad de los registros condiciona las operaciones de identificación, selección, extracción y descarga de datos así como su tabulación y cuantificación. Las posibles imprecisiones en las mediciones bibliométricas, originadas por los errores en los datos suministrados por las BD, han sido descritas prolija y repetidamente (MacRoberts & MacRoberts 1989; Moed & Viriens 1989; Rice et al. 1989; Lardy & Herzhaft 1992).

El nombre de los autores, utilizado frecuentemente para la recuperación de información, es precisamente uno de los campos de los registros en los que se han detectado mayores deficiencias y falta de uniformidad. El escaso o inexistente tratamiento normalizador que presentan, continúa siendo uno de los problemas que con mayor insistencia se viene tratando en la literatura científica de distintas áreas de conocimiento y de distintos sectores que intervienen en el proceso de comunicación y evaluación de la ciencia.

Pero este problema está presente incluso en las BD catalográficas, y ello, pese a que aquí, la elaboración de los registros está sometida a un escrupuloso procedimiento basado en rígidos estándares normativos consensuados internacionalmente. Para la información contenida en los campos de autor que después va a ser utilizada como punto de acceso para la recuperación de información, este procedimiento normalizador se denomina control de autoridades en los nombres personales, cuyo mecanismo básico, que no el único, consiste en adoptar una forma única para un nombre que puede aparecer expresado de formas diferentes en distintos documentos fuente. Su finalidad no es otra que la de asegurar la coherencia y continuidad en su uso para reunir, en el momento de la recuperación, todos los trabajos vinculados a un autor bajo el mismo punto de acceso.

Pues bien, en el caso de los nombres personales españoles que son indizados en las BD, mayoritariamente dominadas por la lengua inglesa, los problemas descritos, tienen, casi con toda seguridad, una mayor incidencia. Aunque en este estudio pretendemos conocer algunas de sus causas, intuimos de entrada, que la complejidad de las estructuras nominales hispanas y sus posibles combinaciones, puestas de manifiesto por el ya lejano trabajo de Taft (1970), unido a la incorrecta interpretación que de las mismas se hace al utilizar criterios de indización propios de las estructuras anglosajonas, están en la base del problema. Es decir, la idiosincrasia de cada país para denominar a las personas, resultado de sus tradiciones históricas, culturales y de uso, es una de las razones que puede explicar la variabilidad de los nombres personales cuando son tratados por sistemas de información lingüísticamente diferentes (Borgman & Siegfried 1992).

En este trabajo, a partir de una muestra de los artículos publicados por autores españoles entre 1987-1996 y recogidos en las BD del Science Citation Index (SCI), de Medline y del Índice Médico Español (IME), los objetivos propuestos son: 1) cuantificación estadística de la magnitud del problema del tratamiento de los nombres españoles, 2) determinar el grado de variabilidad de los

nombres en las tres BD, y 3) valorar esta variabilidad en términos de calidad y coherencia en las prácticas de indización y en términos de recuperación de información por autores nombre personal.

2. MATERIAL Y MÉTODOS

2.1 Muestra

De una población de referencia —autores españoles indizados en las BD SCI, Medline e IME entre 1987 y 1996—, se ha extraído una muestra conformada por los autores pertenecientes a la Facultad de Medicina de la Universidad de Granada (n=171).

La elección de una muestra no probabilística de carácter intencional u opinático como ésta, se justifica por una razón metodológica fundamental: la necesidad de tener identificados previamente de forma completa los nombres de todos los autores. Esta premisa es básica puesto que nos planteamos conocer la variabilidad con que cada uno de estos nombres ha sido indizados. Trabajar con una muestra aleatoria de la población de autores españoles indizados en dichas BD, de procedencia institucional y geográfica muy dispersa, haría imposible, o al menos muy difícil, obtener sus representaciones completas —nombre de pila— de modo preciso.

Por otra parte, lo que realmente importa en este caso para que la muestra sea representativa y con absoluta validez externa de los resultados, es el cumplimiento de los siguientes requisitos:

1. Que en la muestra estén representadas las estructuras básicas imperantes en los nombres personales modernos de autores científicos españoles.
2. Los autores con más de un trabajo deben ser suficientemente representativos para que sus nombres, teóricamente, puedan producir distintas variantes. De los 171 autores recogidos en la muestra, el 93,2% posee más de un trabajo.
3. Para determinadas comparaciones entre las BD, el número de autores que se encuentren indizados en las tres BD del estudio debe ser elevado. Prácticamente los dos tercios de los autores responden a este perfil.

2.2 Análisis de datos

La identificación de los nombres de pila completos de los autores se ha realizado a partir de un listado generado desde las BD que mantiene el Vicerrectorado de Ordenación Académica de la Universidad de Granada, donde figuran los datos personales referidos a todos los profesores de dicha universidad.

Para el tratamiento de los datos, se establece la estructura normalizada de cada autor según los usos catalográficos españoles. Todas las estructuras posibles, de forma total o parcial,

estarían contempladas en la siguiente representación y en las derivadas de los particulares tratamientos, que según las Reglas de Catalogación (RCE 1995), reciben las partículas, que en su caso, pueden aparecer ligando estas unidades:

1º APELLIDO (S Ó C) + 2º APELLIDO (S Ó C) + NOMBRE (S Ó C)

(S ó C= Simple ó Compuesto)

Debajo de la estructura normalizada de cada autor, se listaron las variantes y ocurrencias de las mismas que cada BD ha generado. El criterio general seguido para contabilizar las variantes de cada autor ha consistido en localizar, en cualquier orden en que aparezcan, las coincidencias entre el nombre con que aparece en las BD y la estructura normalizada de base. Para los casos específicos que se citan, se han adoptado las siguientes soluciones:

- Se han considerado variantes también aquellas formas que son producto de un error mecanográfico. Los autores que presentan esta circunstancia están señalados con el signo +.
- En los casos en que la *ñ* es sustituida por la *n* (SCI y Medline), no se considera como una variante distinta.
- Aquellas variantes (sólo cuatro casos) en las que existe la más mínima duda sobre su posible adscripción a uno u otro autor (p.e. variación de un solo carácter) han sido indicados con un *.
- Las variantes producto de la capitalización o no de alguna de las unidades que conforman la estructura normalizada base, han sido contabilizadas como tales variantes.

3. RESULTADOS

3.1 Variantes en la forma de los nombres. Cuantificación

La primera aproximación al problema nos exige una valoración estadística de su magnitud en términos absolutos y relativos. Para ello, hemos cuantificado el número de autores indizados en las bases de datos estudiadas distribuidos por el número de variantes que presentan en su forma del nombre (Tabla 1).

Tabla 1

Número de autores con n variantes en la forma del nombre por BD

Variantes (Formas)	**SCI**		**MEDLINE**		**IME**	
	Nº Autores	%	Nº Autores	%	Nº Autores	%
1	65	51.59	63	47.73	50	33.33
2	37	29.36	43	32.57	78	52
3	23	18.25	14	10.61	12	8
4	1	0.79	10	7.57	8	5.33
5	0	0	1	0.76	1	0.66
6	0	0	1	0.76	0	0
7	0	0	0	0	1	0.66
+7	0	0	0	0	0	0
Total	126	100	132	100	150	100

Sí como hemos señalado en la introducción, la unificación de los puntos de acceso —nombre de autor en este caso— es un principio básico para asegurar la calidad de la recuperación de información en las bases de datos, la tabla 1 muestra, que la proporción de autores que se ven afectados por el incumplimiento de este principio (2 ó + variantes) es muy significativa, con porcentajes entorno al 50% en el caso de SCI y MEDLINE y con el 67% en el caso de la base de datos española IME, en la que teóricamente, por el mejor conocimiento que se tiene de las estructuras nominales españolas, la incidencia debería ser menor.

La lectura inicial que se puede hacer de estos datos es la siguiente: si en un proceso de búsqueda por nombre personal pretendemos recuperar todos los trabajos de un conjunto determinado de autores, para el 50% de ellos, tendríamos que efectuar dos o más intentos de búsqueda, o bien, utilizando las técnicas de browsing a partir de los índices, tendríamos que intuir que trabajos representados por formas distintas de un mismo nombre, algunas de ellas situadas próximas en la lista, pertenecen a un mismo autor.

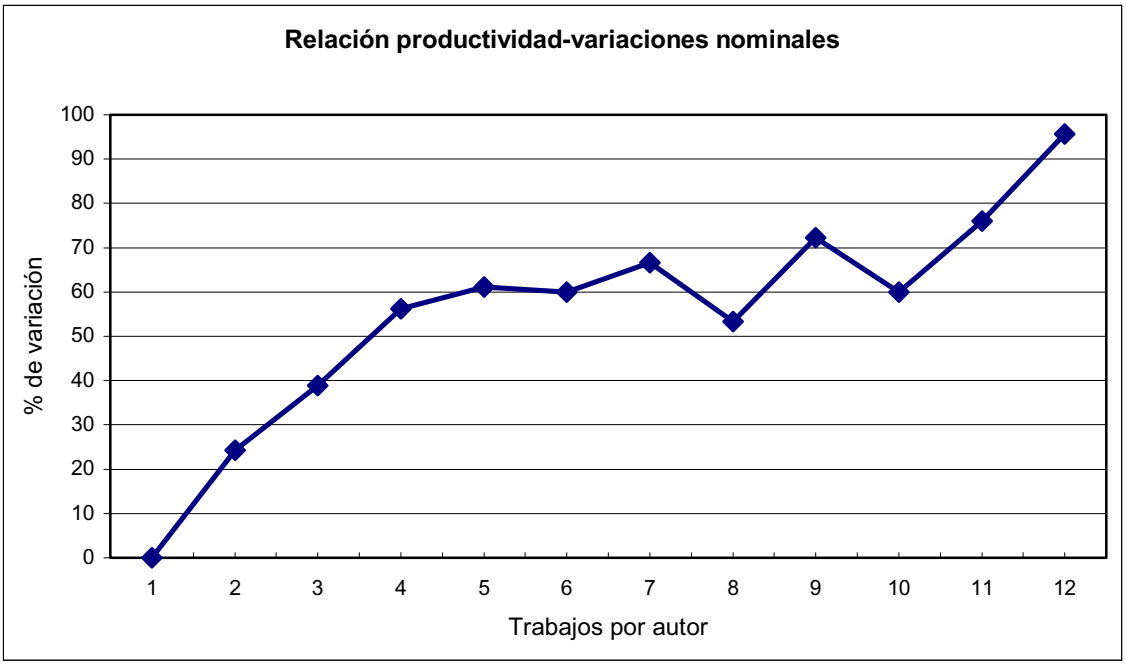
Por otra parte, puesto que no existen razones lógicas que nos indiquen que el principio de unificación y control pudiera estar aplicándose de forma indiscriminada en las BD a determinados autores, la explicación del también considerable número de autores asociados con una variante (el 50% en SCI y Medline, y el 33.3% en IME) solo podemos encontrar incorporando al análisis la variable de productividad. Esto quiere decir, que los autores con una sola variante obedecen a alguna de las razones siguientes: o bien están representados en la BD por un solo registro (trabajo), o bien, siendo un autor con más de 1 registro ha adoptado un nombre de pluma uniforme al firmar sus trabajos.

Pero además, la introducción de la variable productividad, no sólo arroja información sobre este particular, sino que nos descubre, y esto es lo más importante, interesantes interpretaciones individuales y comparativas sobre calidad de las BD en cuanto a sus prestaciones de recuperación de información.

3.2 Relación entre productividad y variantes

Con el fin de obtener una cualificación más detallada del tratamiento que reciben los autores españoles, las variaciones en los nombres han sido distribuidas en función de la productividad de los autores. Para su análisis, los datos han sido agrupados bajo las dos formas que, a nuestro juicio, muestran los aspectos más relevantes del problema. En la primera de ellas, con los datos acumulados de las tres bases de datos, la representación de la figura 1 nos muestra la tendencia que sigue relación existente entre el aumento de la productividad de los autores y el aumento de las variaciones en sus nombres en términos porcentuales.

Figura 1



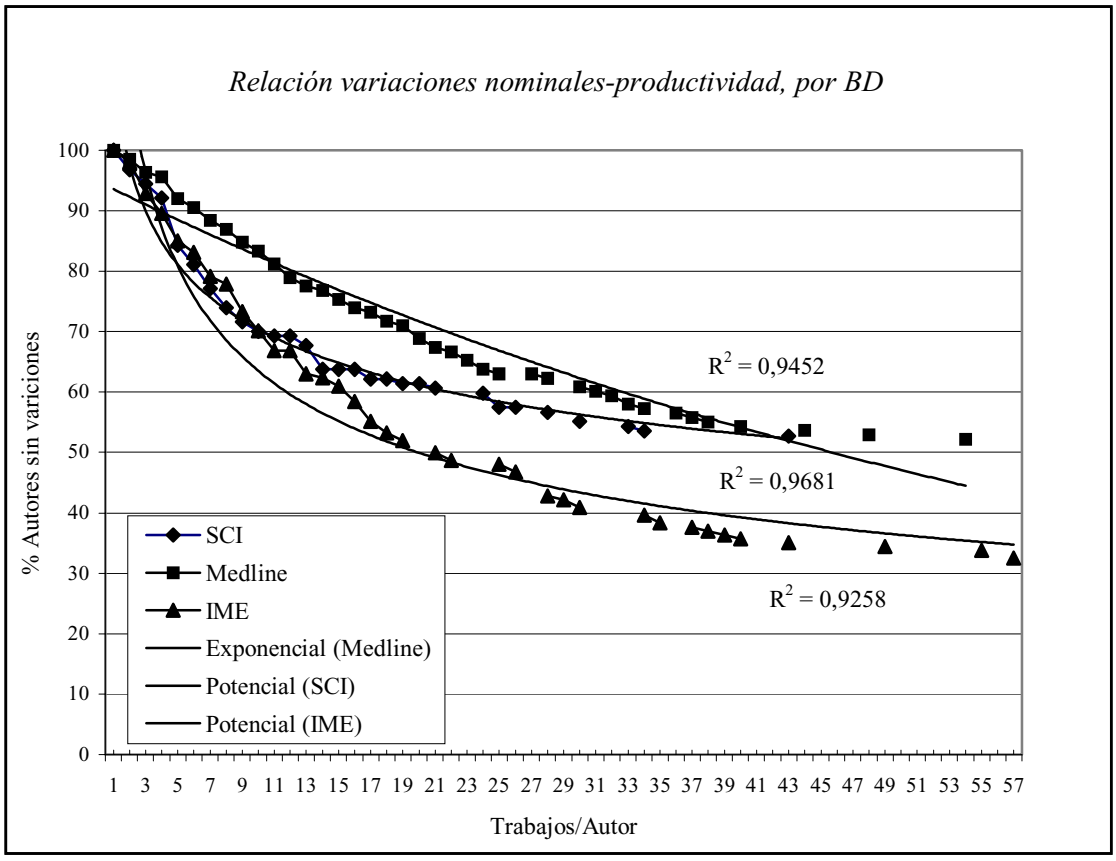
Considerando que el % de variación es 0 para aquellos autores con un solo trabajo, a medida que aumenta la producción de los autores, aumenta el número de variantes con las que aparecen identificados sus nombres en las siguientes proporciones: para autores cuya productividad está comprendida entre 2 y 3 trabajos, las variaciones afectan a una banda comprendida entre el 25 y el 40% de los autores. Para los autores comprendidos entre 5 y 10 trabajos, que podríamos calificar como productivos, las variaciones afectan ya a más del 50% de los autores, y en algunos casos al 70% (9 trabajos). Entre los autores que podríamos clasificar como muy productivos (más de 10 trabajos en el período estudiado), la variabilidad, por encima del 70%, llega a acercarse al 100% de los casos.

En términos de recuperación de información la consecuencia es, que la identificación y localización de los trabajos producidos por un autor, si se realiza mediante una búsqueda simple por sus apellidos, seguiría una curva de rendimiento decreciente inversa a la representada en la figura 1, tendente a 0 en los casos de los autores muy productivos. Por otra parte, y dado que por ahora

estamos tratando los datos independientemente de sí la forma indizada del autor es la más adecuada o no para los nombres españoles, los autores con bajas frecuencias de producción obtendrían un mayor rendimiento en las búsquedas si asumimos que el autor fuera buscado por la forma utilizada por las BD, cosa que no tiene porque suceder.

En la segunda representación (figura 2) se presentan los resultados desagregados por Bases de Datos. Aquí, la relación productividad-variaciones en los nombres de los autores se establece en sentido descendente, es decir, vemos como a medida que se incrementa el número de trabajos por autor, disminuye el porcentaje de autores sin variación, esto es, disminuye la uniformidad en la forma del nombre.

Figura 2



Visto sobre líneas de tendencia y en términos de efectividad en la recuperación de información, el aumento en la producción de los autores significa una disminución de esta efectividad, que se adapta a un modelo potencial negativo (en dos de las BD) del volumen de trabajos teóricamente recuperables, que, en el caso de SCI alcanza el 50%, y, en el caso de IME, el 70%. Es decir, del total de los registros potencialmente recuperables correspondientes a los trabajos publicados por los investigadores médicos de la muestra y recogidos en estas BD, entre el 50 y 70% de los trabajos aparecen indizados por diferentes formas del nombre, cualquiera que esta sea. El caso es especialmente dramático para IME, base que por su origen nacional debería presentar una mayor unificación, aunque sólo fuera por el mayor conocimiento que sus gestores tienen de las

estructuras de nombres españoles, independientemente de sí se llegan a aplicar o no criterios normativos de referencia.

Por otro lado es interesante señalar que Medline presenta un descenso que se ajusta a un modelo diferente, exponencial, el menos acentuado de las tres bases, al menos hasta la frecuencia 20. Este descenso más suave está relacionado probablemente con un control básico de los nombres. No olvidemos que su producción está a cargo de la National Library of Medicine.

4. DISCUSIÓN Y CONCLUSIONES

Aunque el estudio necesita una mayor profundización, y de momento los resultados son solo aplicables a los autores de medicina, si podemos obtener algunas conclusiones generales. La magnitud del problema descrito obedece, fundamentalmente, a las siguientes causas: 1) las variaciones en la forma que a lo largo de su producción científica viene utilizando un autor para firmar sus trabajos, y que quedan reflejadas en las fuentes (revistas) de las que emanan las referencias bibliográficas, concretamente en los campos de autor, 2) falta de rigor y homogeneidad en los procesos de tratamiento (control de autoridades) de la información por las BD, 3) individualmente, las BD aplican determinados procedimientos de indización, pero escasamente adecuados al tratamiento de las estructuras nominales españolas, y muy sesgadas hacia las estructuras lingüísticas inglesas, precisamente por el dominio de las bases de datos anglosajonas, 4) el conocimiento del problema por parte de algunos autores y las precauciones adaptadas consiguen medidas correctoras muy pobres y no siempre con resultados satisfactorios desde el punto de vista de la normalización de los nombres españoles.

Por todo ello, es preciso iniciar un proceso de actuación tendente a emitir recomendaciones a nuestros autores para que extremen su atención al presentar sus trabajos; a las revistas para que apliquen las normas internacionales y a las bases de datos para que establezcan mecanismos de control y unificación.

REFERENCIAS

BELL, J., SPEER, S. (1988). Bibliographic verification for interlibrary loan: is it necessary?. *College & Research Libraries*, 49: 494-500.

HAWKINS, D.T. (1977). Unconventional uses of on-line information retrieval systems: on-line bibliometric studies. *Journal of the American Society for Information Science*, 28(1): 13-18.

MOED, H.F., VIRIENS, M. (1989). Possible inaccuracies occurring in citation analysis. *Journal of Information Science,* 15: 95-117.

RICE, R.E., BORGMAN, C.L., BEDNARSKI, D., HART, P. J. (1989). Journal-to-journal citation data : issues of validity and reliability. *Scientometrics*, 15(3-4): 257-282.

LARDY, J.P, HERZHAFT, L. (1992). Bibliometric treatments according to bibliographic errors and data heterogeneity: the end-user point of view. *Online Information 92, Proceedings of the 16th International Online Information Meeting, London, 8-10 December 1992. Edited by David I. Raitt, Oxford and New Jersey, Learned Information*.

TAFT, R.L. (1970). Name search techniques. Bureau of Sytem Development, New York State Identification and Intelligence System, Alcany, NY (Special Rep. Nº 1).

BORGMAN, C.L., SIEGFRIED, S.L. (1992). Getty's Synoname and its cousins: a survey of applications of personal name-matching algorithms. *Journal of the American Society for Information Science*, 43(7): 459-476.

RCE 1995. Reglas de catalogación. Ed. refundida y rev. Madrid: Dirección General del Libro, Archivos y Bibliotecas, 1995